Supplementary Information for

Signal Peptides Generated by Attention-Based Neural Networks

Zachary Wu[†,1], Kevin K. Yang[†,2], Michael J. Liszka[†,3], Alycia Lee[4], Alina Batzilla[5], David Wernick[3], David P. Weiner[3], Frances H. Arnold[*,1]

[1] Department of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA, USA

[2] Generate Biomedicines, Cambridge, MA, USA; work performed while at California Institute of Technology

[3] BASF Enzymes, San Diego, CA, USA

[4] Department of Computational and Mathematical Sciences, California Institute of Technology, Pasadena, CA, USA

[5] Department of Molecular Biotechnology, Heidelberg University, Heidelberg, Germany; work performed while at BASF Enzymes

*To whom correspondence may be addressed. Email: frances@cheme.caltech.edu

Tables

Supplemental Sections

Supplemental Table 1: Primers used to generate linear DNA fragments

Primer	Sequence
pHT vector fwd	CAGACTTTCTAGAGGTCTCATAGCGCAGCC
pHT vector rev	TGAGCTCCTCGAGGGTCTCTATATAGAGTC
Signal Peptide fwd	TCCGCCTGACCTCCATGGGGTCTCAATATG
Signal peptide rev	CTCCACCTAGCCTGATATCGGTCTCTAGCC
Gene of Interest fwd	ATTCCGCCTGACCTGGTCTCAGGCTG
Gene of Interest rev	CTGAGCCTCCACCTAGCCTGGTCTCTGCTA

Supplemental Table 2: Enzyme activity reaction conditions

Enzyme Class	Substrate	Buffer	Reaction Conditions
Amylase	Red Starch (Megazyme)	50mM NaCl, 50mM BTP, 50mM Citric acid, 50mM N-cyclohexyl-3-aminopropanesulfonic acid (CAPS), pH 7.0	20ul of the cell supernatant were added to 10ul of 2% (w/v) Red Starch substrate in buffer and the mixture was incubated at 40 °C for 30 min. Reactions were quenched with 50ul of 95% ethanol and the assay plate was centrifuged at 4000 rpm for 10 min. The absorbance was measured in 60ul of the supernatant solution at 510nm.
Lipase	C8-PNP (4-Nitrophenyl octanoate, Sigma)	100mM HEPES, pH 7.5, 100mM NaCl, 20mM $CaCl_2$, 0.1% Triton-100	10ul of the cell supernatant were added to 90ul of 0.8 mM C8-PNP substrate in buffer. The absorbance was measured at 405nm kinetically for 15 min at a 31 sec interval. The initial reaction rate was calculated from the linear portion of the kinetic read.
Protease	AAPF (N-Succinyl-Ala-Ala-Pro-Phe p-nitroanilide, Sigma)	100mM HEPES buffer, pH 7.5	10ul of the cell supernatant were added to 90ul of 2mM AAPF substrate in buffer. The absorbance was measured at 410nm kinetically for 15 min at a 31 sec interval. The initial reaction rate was calculated from the linear portion of the kinetic read.
Xylanase	Azo-Xylan Birchwood (Megazyme)	50mM NaCl, 50mM Bis-tris propane (BTP), 50mM Citric acid, 50mM N-cyclohexyl-3-aminopropanesulfonic acid (CAPS), pH 7.0	20ul of the cell supernatant were added to 20ul of 2% (w/v) Azo-Xylan substrate in buffer and was incubated at 40 °C for 30 min. Reactions were quenched with 100ul of 95% Ethanol and the assay plate was centrifuged at 4000 rpm for 10 min. The absorbance was measured in 90ul of the supernatant solution at 590nm.

Supplemental Table 3: Enzyme activity reaction conditions

Strain	Modifications	Reference
Bacillus subtilis PY79 base strain	ΔnprE ΔaprE Δepr Δmpr ΔnprB Δvpr Δbpr ΔsigF ΔskfA Δxpf ΔlytC ΔsdpC	Provided by BASF
BS – Prot	PY79 base strain ΔskfA ΔspbC ΔskfA	Provided by BASF
BS – Amy	PY79 base strain ΔskfA ΔspbC ΔskfA ΔAmyE	Provided by BASF
BS – Xyl	PY79 base strain ΔskfA ΔspbC ΔskfA ΔxynA ΔxynC	Provided by BASF
BS – Lip	PY79 base strain ΔskfA ΔspbC ΔskfA ΔestA ΔestB	Provided by BASF

Supplemental Table 4: Control sequences generated by the trained Transformer for random inputs.

	Randomized Inputs	Generated SPs
Bacillus AA distribution	DTTRFTAIEIFTTHWSGMVLFMSIIIVVFFSIGDSGVGDANIFYHQARLFKA DKIANVQTLKGVYSNPLDYYFNHDLPRSGRLNLGNSHHIQLNKLQILF	MKLLVLLILILIPVVAWA
Bacterial AA distribution	SEVHPENTYQEGLPGIKALRIQFMTATIVHEGSTIDEVVRKKALTQTTCIPY LIAASIIVSSGEDMHIIVIKHMAQQVLPKLISCAMEHFVGPSHAEASE	MNWLVIMFAIGLACSTSLA
Uniform AA distribution	DIVPSLYPPAKSDERKVELTSKQSELNSENLKATQQILKDGYATPGRVRKID AETKKQLNLIKHTYRTVRKQEDKGVYQASNPRNFKFIVPQLVQYKGRN	MEHTLRILVCSLLFTVAIWG

Supplemental Table 5: Distribution of Protein and SP lengths, as obtained from UniProt.

Length (Amino Acids)	Signal Peptide	Protein
Mean ± StDev	23.5 ± 6.2	422.5 ± 454.1
Minimum	10	21
Median	22	323
Maximum	69	13076

Supp Section I: Comparing various generation approaches

In this section, we compare four methods of generating SP sequences for their efficacy: (A) a profile Hidden Markov Model (pHMM), (B) a heuristic-based generation approach, (C) a variational autoencoder (VAE)[1] trained on just SP sequences, and (D) the Transformer-based approach. In the first round of comparison, we query the SignalP 5.0 server with these SP sequences prepended to protein sequences to obtain SignalP's probability of functioning.

(A) pHMM: We first generated an alignment with Clustal Omega[2] of SPs from our training. We used 4000 randomly selected SPs, as Clustal Omega is limited to 4000 sequences. From this point on, we use default settings in S. Eddy's HMMER package.[3] From the Clustal Omega alignment, we then use HMMER's *hmmbuild* to build a profile HMM. From this profile HMM, we use HMMER's *hmmemit* to generate protein samples. As HMMs typically follow the first order Markov property, the HMMs do not have a concept of length, and can generate longer sequences. We generate 1024 constructs by emitting 1024 SP sequences from *hmmemit*, each of which is prepended to one of the 41 enzymes tested.

(B) Heuristics: We follow the following heuristics from a recent review by Owji, Hajar, et al.[4] The N-region (1-6 residues) contains at least 1 positive residue (KR). The H-region (7-15 residues) contains hydrophobic residues (AILMFWYV). The C Region (3-7 residues) contains uncharged residues (STNQAVILMFYW) followed by an AXA motif, where X is not in C/P/A. 768 SPs were generated by heuristics and randomly matched to one of the 41 input proteins.

(C) VAE: We train a VAE with fully connected layers and ReLU activation on SP sequences. The sizes of the hidden layers are 400 and 20 each for the mean and variance of the latent encoding. We then randomly sample 1024 SP sequences, each of which is prepended to one of the 41 enzymes tested.

(D) Transformer: As described in the main text.

We then query these constructs with SignalP v5.0 to obtain the probability of containing a Signal Peptide in gram-positive hosts.[5] The results are shown visually in the figure below, and summarized in its accompanying table.

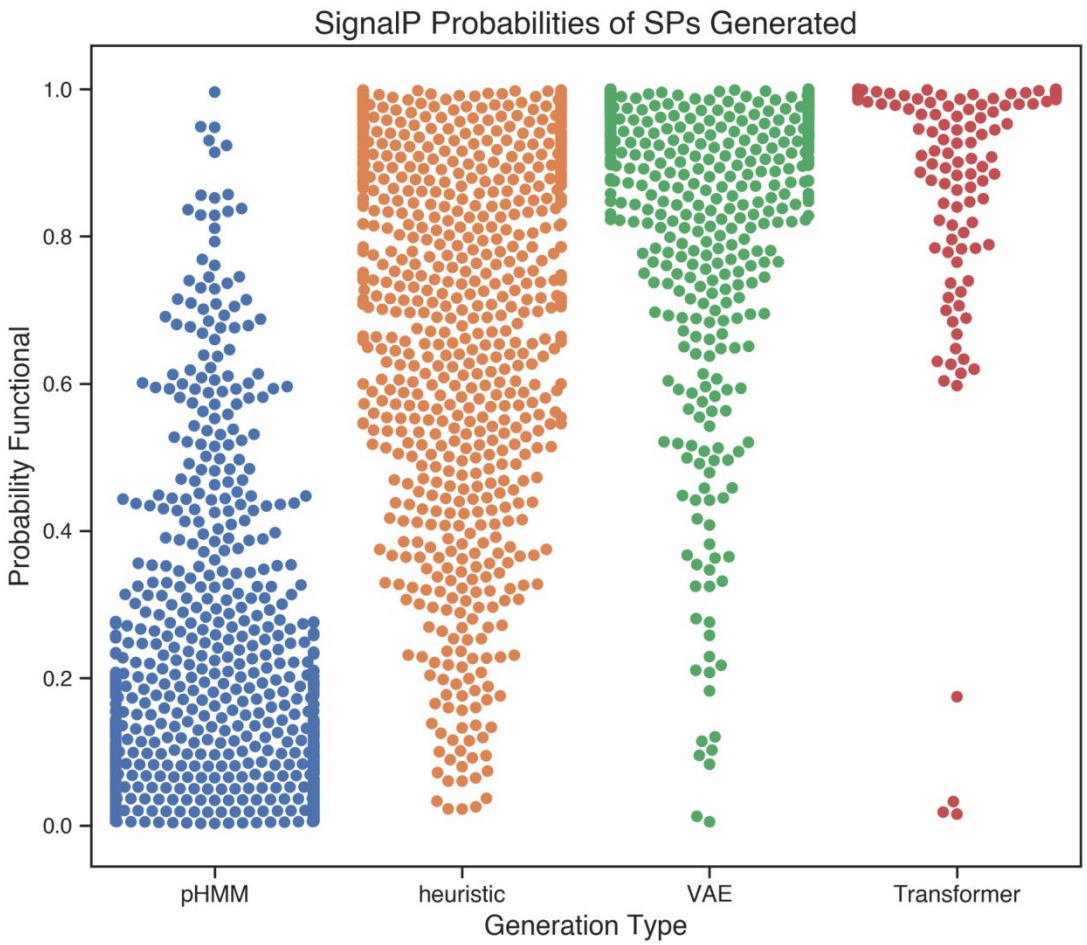


Figure S1: Probability of generated sequences as predicted by SignalP 5.0 The sequences are generated by either a pHMM, heuristic-based approach, VAE, or Transformer model.

Table S1: Summary of probability of predictions for various generation methods, as predicted by SignalP 5.0

Method	Mean SignalP Probability
pHMM	15.6% ± 19.2%
heuristic	70.8% ± 25.3%
VAE	**92.4% ± 15.2%**
Transformer	90.4% ± 17.1%

The VAE and Transformer significantly outperform the other two methods (p-value < 10^{-20}), but are not statistically different from each other (p-value = 0.251). As both are predicted to function with high probability by SignalP, we turn to other methods of analysis below.

First, we compare the log-likelihood and percent (unaligned) sequence identity on an identical, withheld validation set of 5707 SP sequences for the VAE and Transformer, which are comparable between the two methods, although the VAE is more accurate by this metric.

Method	Log likelihood on Validation Set	% Sequence Identity (Unaligned)
VAE	**-2.86**	**45.4%**
Transformer	-3.01	34.5%

However, in closer examination of the sequences generated by the VAE and Transformer, an issue of repeated residues begins to emerge:

Sample VAE Sequences	Sample Transformer Sequences
MRKRLLALALAALSLLLLLSFGVKALAGSGA	MKFTQAVLSLLGSAATALA
MRLLLLLLLVLLLAAPAPPGLS	MKLKKLGVILAICLGISSTFA
MKLLLLLVTLLTSVLALQA	MRVLSATAFLALLAFGLSGATA
MRLLLLALLAAAAVALASA	MLFKSVLLALASAGVAVNA
MASSSSSLFVVLAVLLLLLLLTLSSA	MNISIFVGKLALAALGSALVA

Namely, the hydrophobic region appears to be particularly saturated in repeating residues for VAE sequences. For a more comprehensive comparison, we determined the average length of substrings with maximum *R* repeating residues (e.g.: an example of *R*=1 is LLLLLLL, *R*=2 is LLLILIL, *R=3* is LALLAI).

	Average longest substring from VAE Seqs	Average longest substring from Transformer Seqs
R = 1	3.9 ± 1.9	2.1 ± 0.6
R = 2	7.2 ± 3.0	4.0 ± 1.0
R = 3	10.7 ± 3.5	6.3 ± 1.5

As biological sequences rarely contain such long repeats, we elect to proceed with the Transformer generated sequences in experimental validation.

Supp Section II: Functionality classification from experimental validation

Classification Explanation

SP-enzyme constructs were generated stochastically and tested in biological replicates. Thus, some random SP-enzyme constructs contain the same SP and enzyme but are not part of the same set of biological replicates.

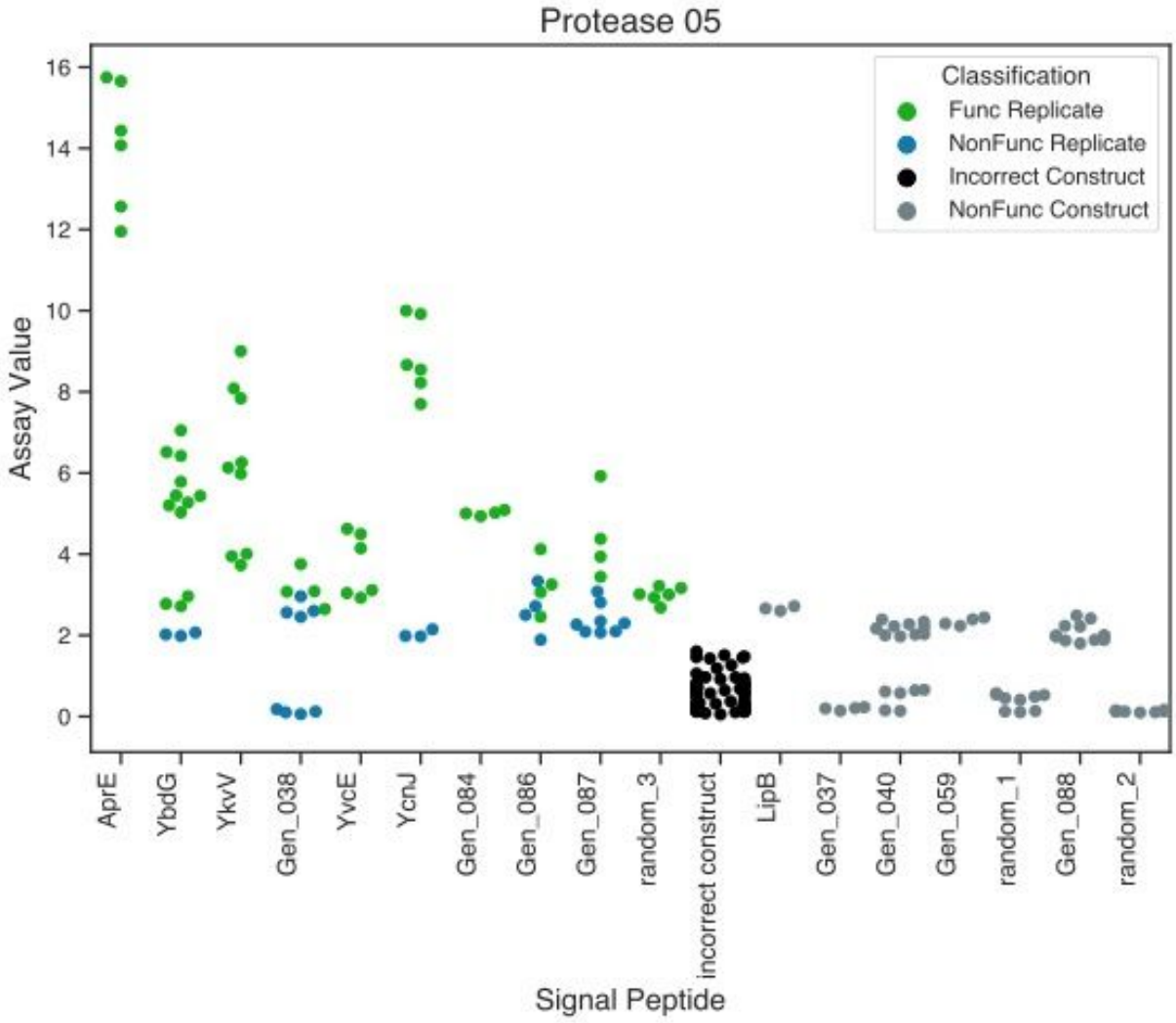
An SP-enzyme construct is classified as functional if any set of biological replicates satisfies two conditions (p-value and effect size):

1) The p-value for the set of replicates compared to the negative controls (constructs without either an SP or enzyme) was less than 0.05 in a two-sided t-test with unequal variance for two independent samples of scores. Biological replicates are grown from the same colony, after which they are processed separately. The null hypothesis is that these two independent samples have identical expected values. The scipy.stats implementation (ttest_ind) was used. (p-value)
2) The difference between the mean of the replicates and the mean of the incorrect constructs is greater than 2 times the difference between the mean of the incorrect constructs and the highest measured activity of the incorrect constructs. (effect size)

The second condition is incorporated to gain some concept of effect size: the *activity* of the SP-enzyme construct must be high, not just statistically significant.

In the following plots, constructs classified as functional are to the left of the "incorrect constructs", which are shown in black. For functional constructs, some sets of replicates (picked from the same colony) would have been classified as nonfunctional. These are shown in blue. The functional sets of replicates are shown in green. Finally, constructs classified as nonfunctional are shown in grey. Summary statistics and notes highlighting important points are also provided.

All generated Signal Peptides can be found in Supplemental File 1.



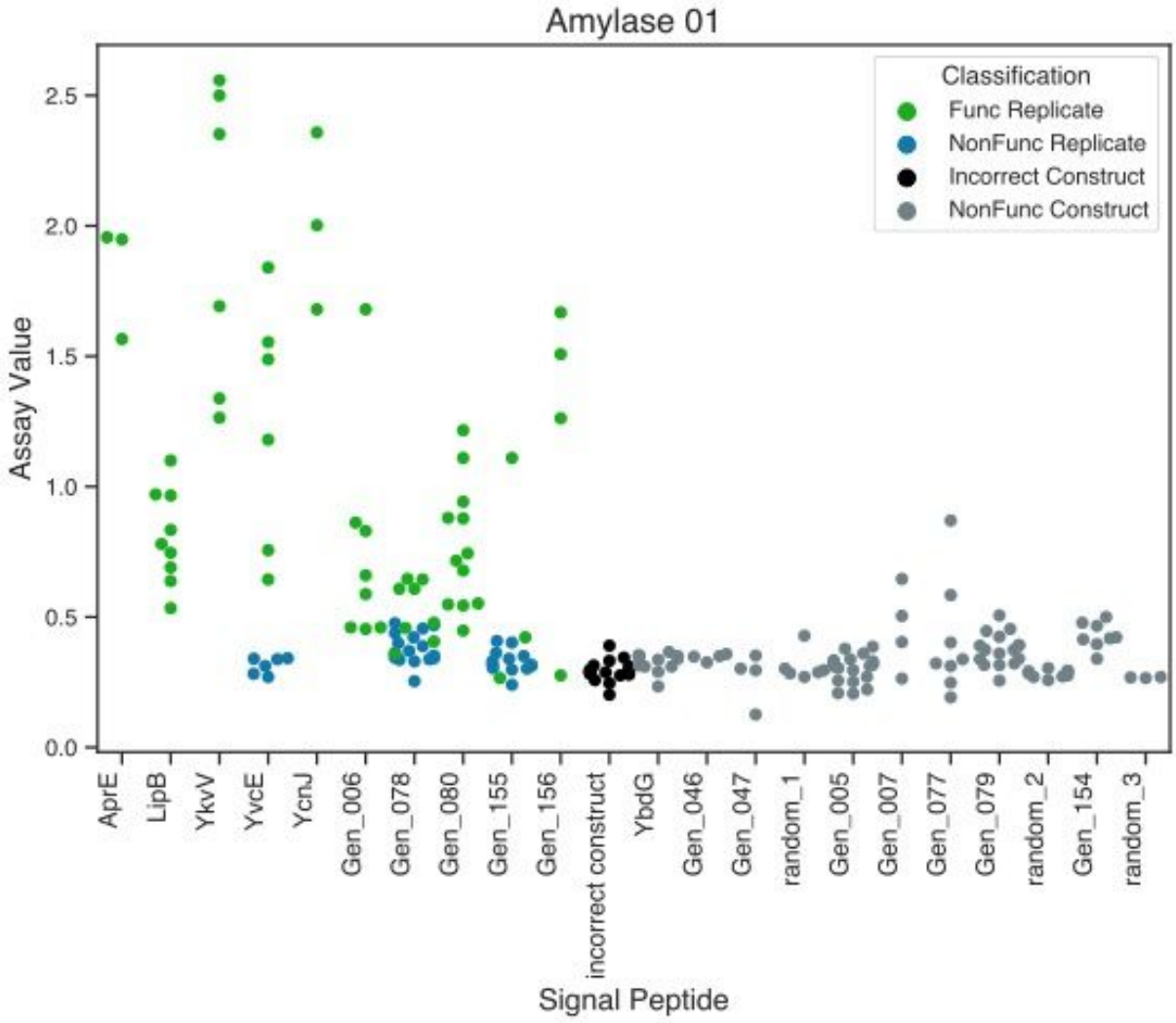
Shuffled_ID: seq1
Original_ID: Arnold_036
Plot_ID: Protease 05

Construct Functionality Classification Summary
0.33 : 1 out of 3 constructs with SPs generated for random inputs functional*
0.83 : 5 out of 6 constructs with natural SPs functional
0.50 : 4 out of 8 constructs with generated SPs functional

Assay information is provided in Supplemental Table 2.

*The SP sequence for the control generated for a randomized input that was classified as functional is

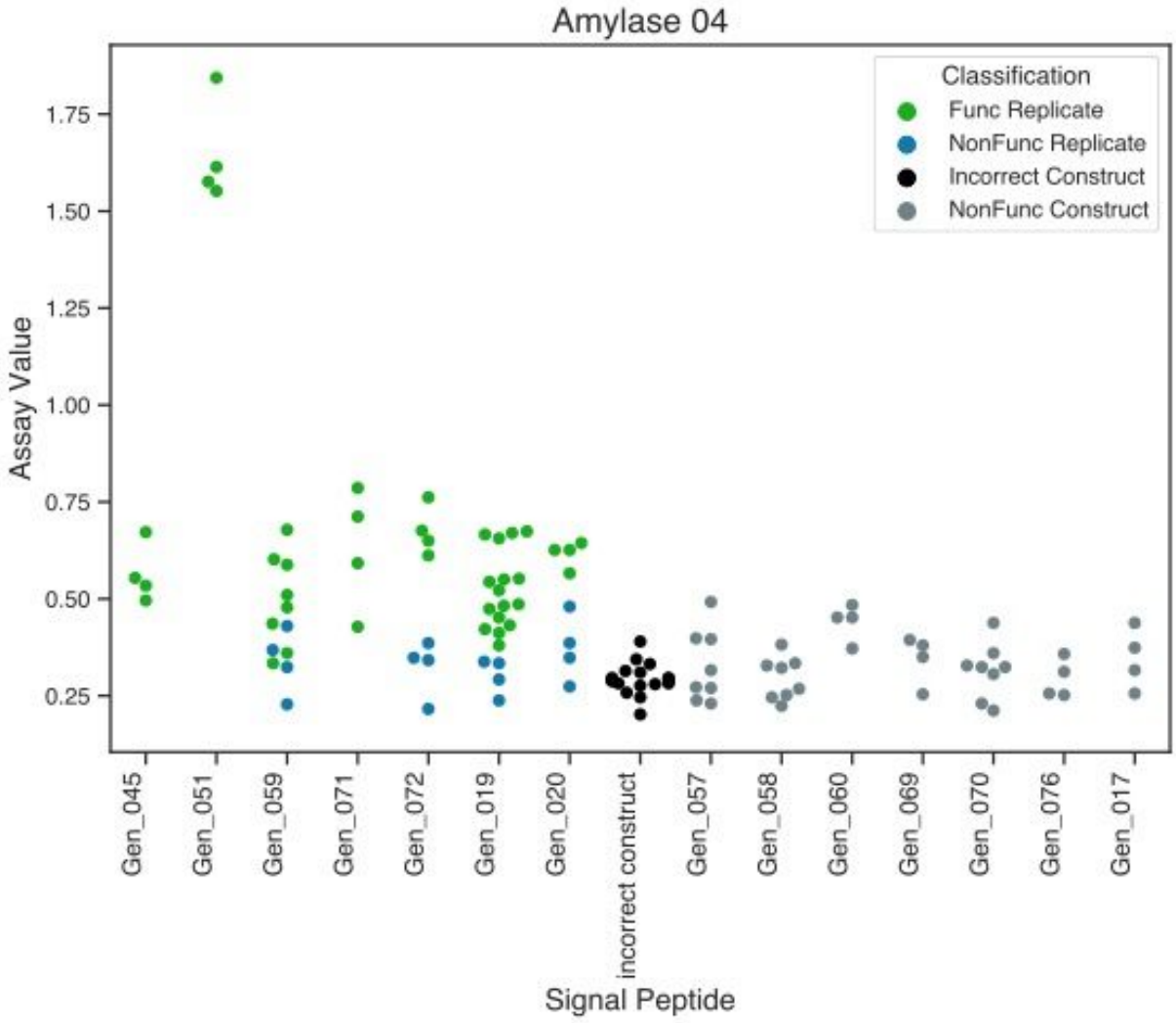
MEHTLRILVCSLLFTVAIWG.



Shuffled_ID: seq39
Original_ID: Arnold_001
Plot_ID: Amylase 01

Construct Functionality Classification Summary
0.00 : 0 out of 3 constructs with SPs generated for random inputs functional
0.83 : 5 out of 6 constructs with natural SPs functional
0.42 : 5 out of 12 constructs with generated SPs functional

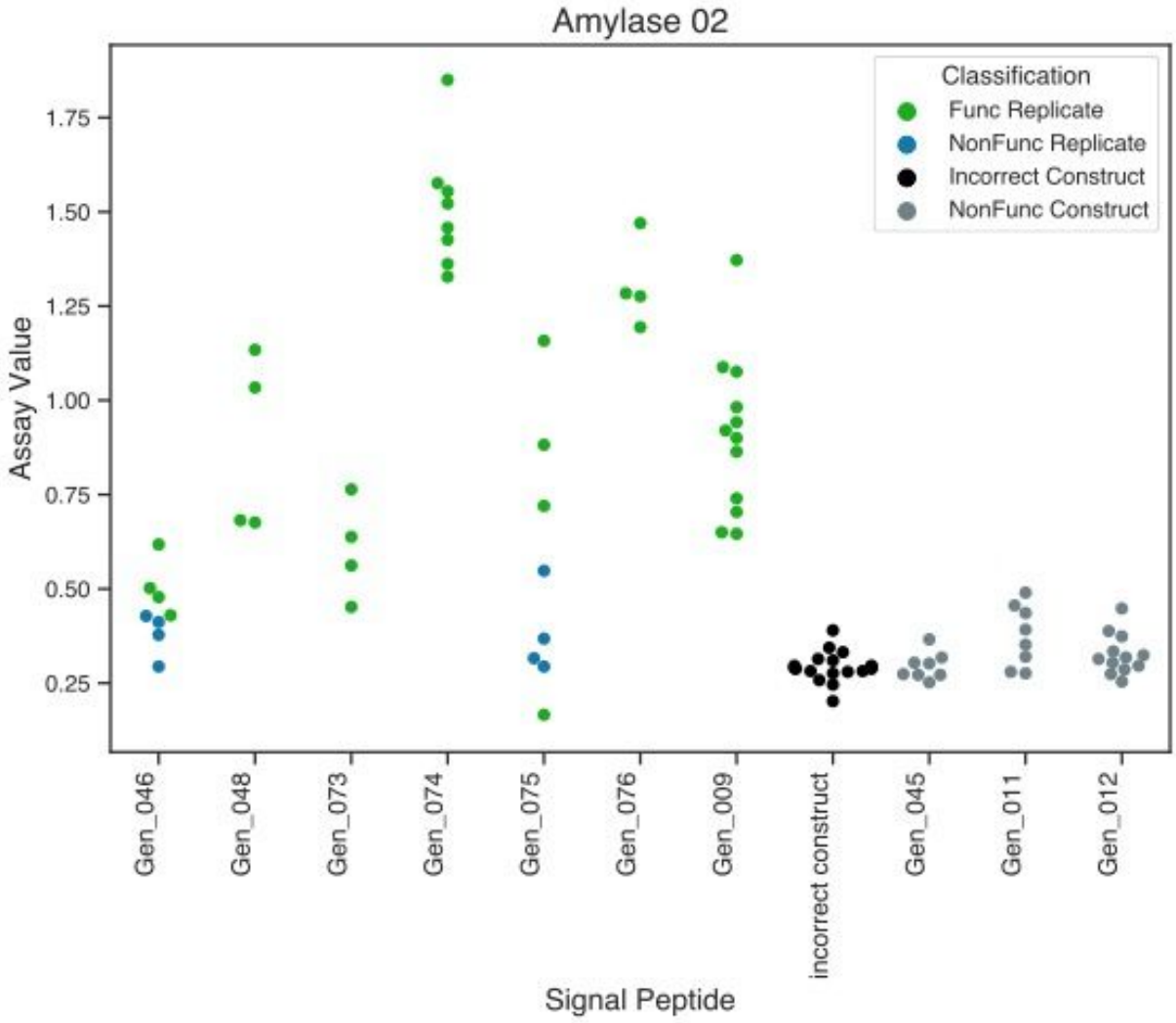
Assay information is provided in Supplemental Table 2.



Shuffled_ID: seq5
Original_ID: Arnold_004
Plot_ID: Amylase 04

Construct Functionality Classification Summary
No constructs with SPs generated for random inputs tested.
No constructs with natural SPs tested.
0.50 : 7 out of 14 constructs with generated SPs functional

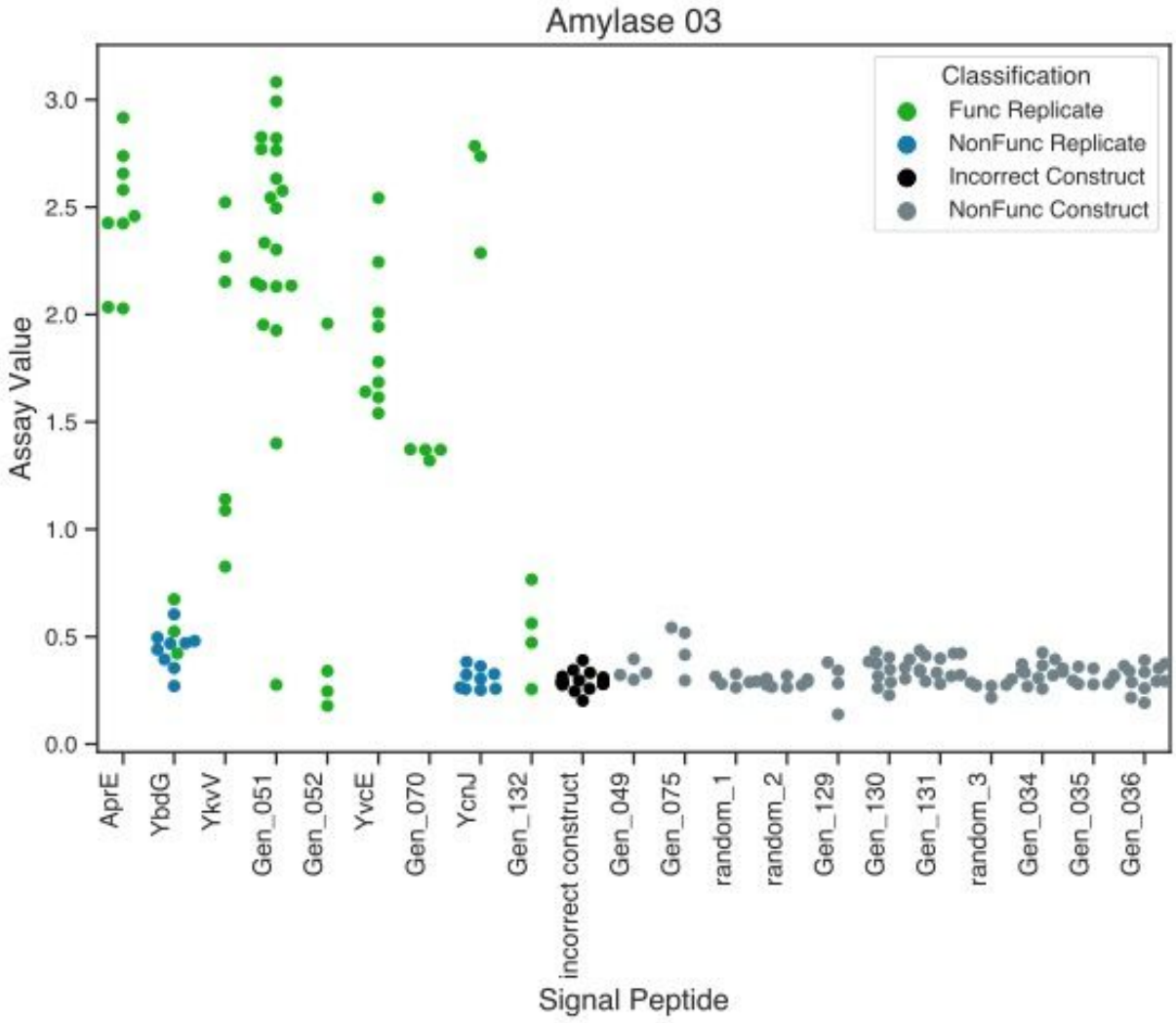
Assay information is provided in Supplemental Table 2.



Shuffled_ID: seq19
Original_ID: Arnold_002
Plot_ID: Amylase 02

Construct Functionality Classification Summary
No constructs with SPs generated for random inputs tested.
No constructs with natural SPs tested.
0.70 : 7 out of 10 constructs with generated SPs functional

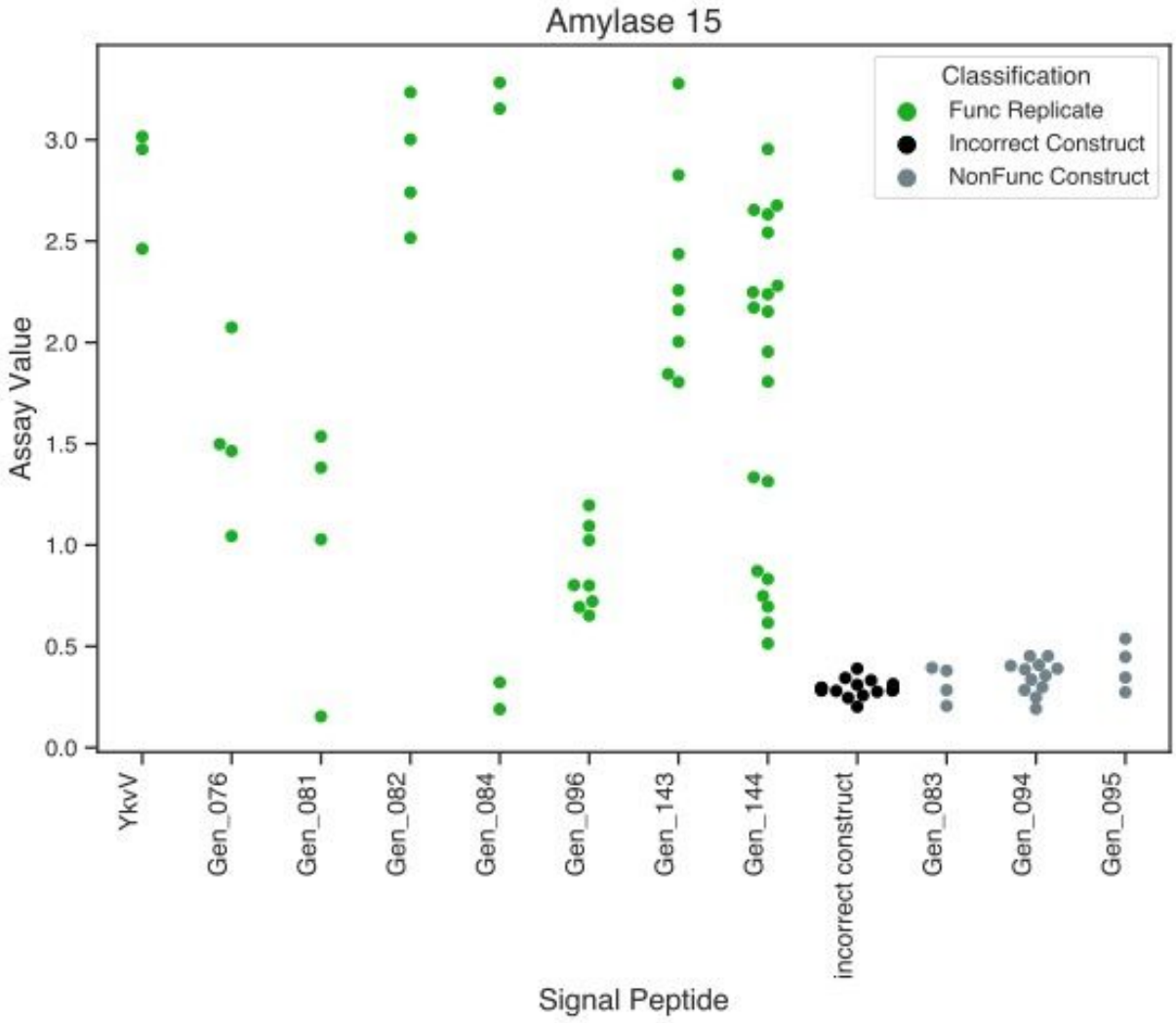
Assay information is provided in Supplemental Table 2.



Shuffled_ID: seq33
Original_ID: Arnold_003
Plot_ID: Amylase 03

Construct Functionality Classification Summary
0.00 : 0 out of 3 constructs with SPs generated for random inputs functional
1.00 : 5 out of 5 constructs with natural SPs functional
0.33 : 4 out of 12 constructs with generated SPs functional

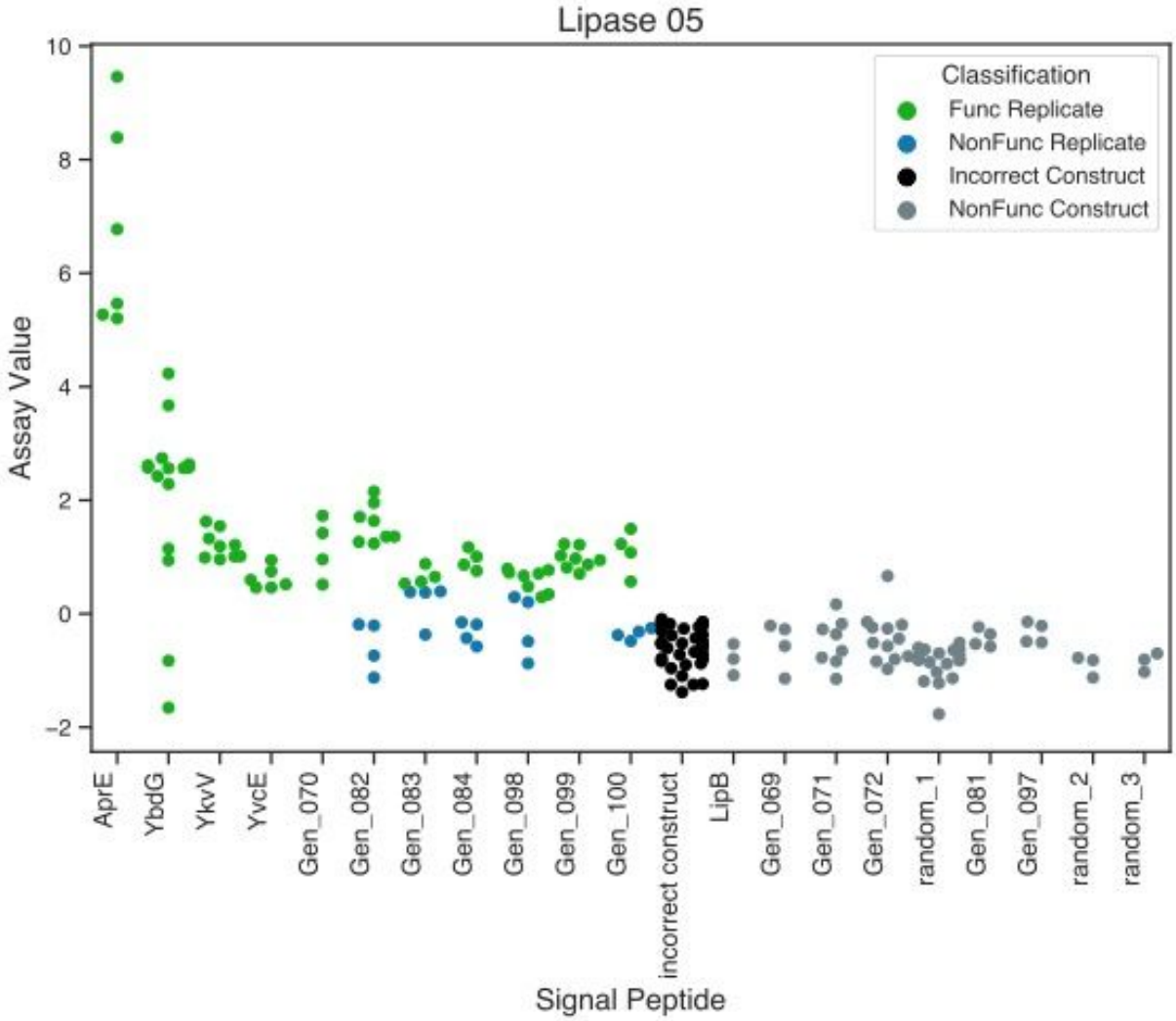
Assay information is provided in Supplemental Table 2.



Shuffled_ID: seq21
Original_ID: Arnold_015
Plot_ID: Amylase 15

Construct Functionality Classification Summary
No constructs with SPs generated for random inputs tested.
1.00 : 1 out of 1 constructs with natural SPs functional
0.70 : 7 out of 10 constructs with generated SPs functional

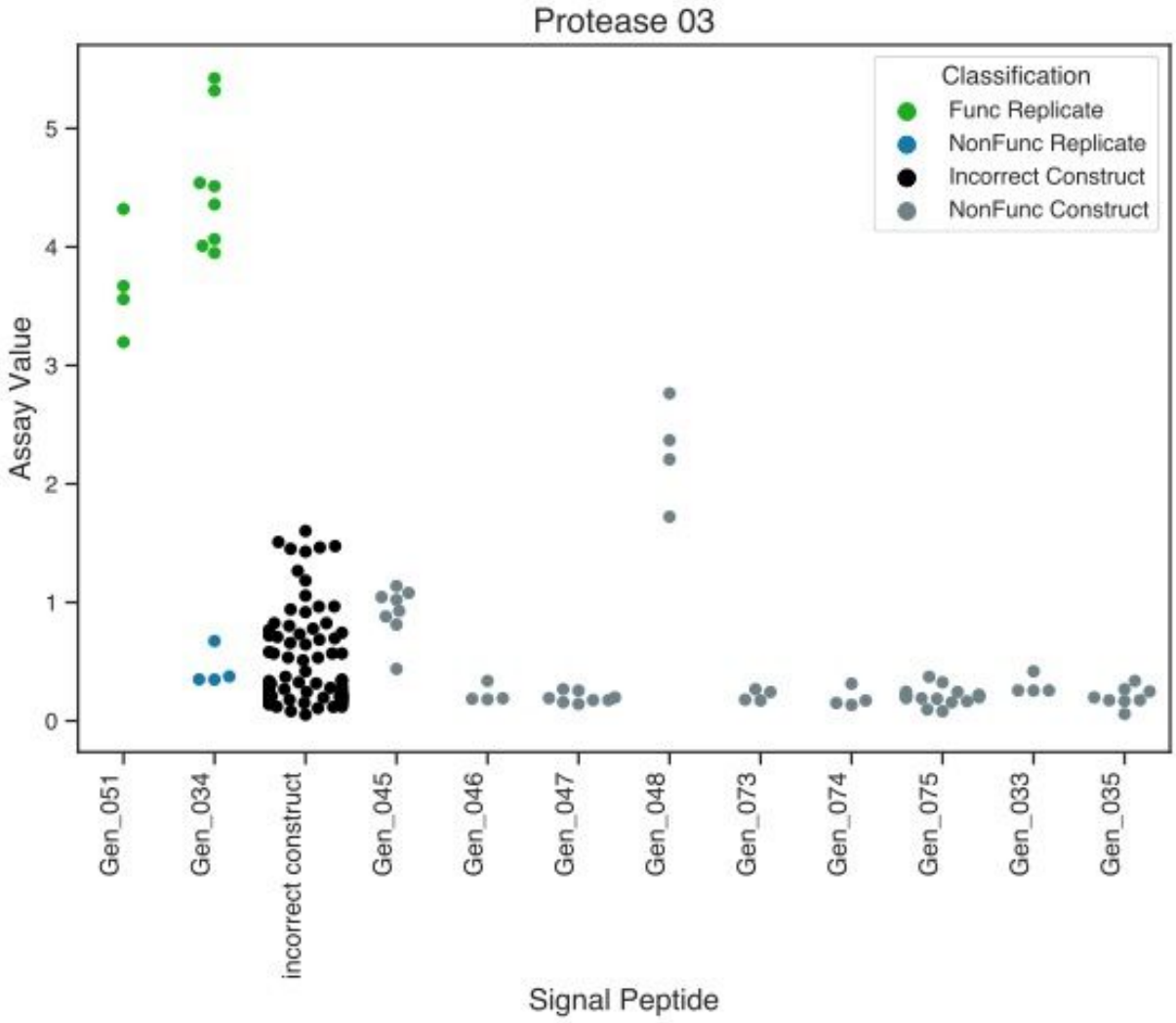
Assay information is provided in Supplemental Table 2.



Shuffled_ID: seq25
Original_ID: Arnold_026
Plot_ID: Lipase 05

Construct Functionality Classification Summary
0.00 : 0 out of 3 constructs with SPs generated for random inputs functional
0.80 : 4 out of 5 constructs with natural SPs functional
0.58 : 7 out of 12 constructs with generated SPs functional

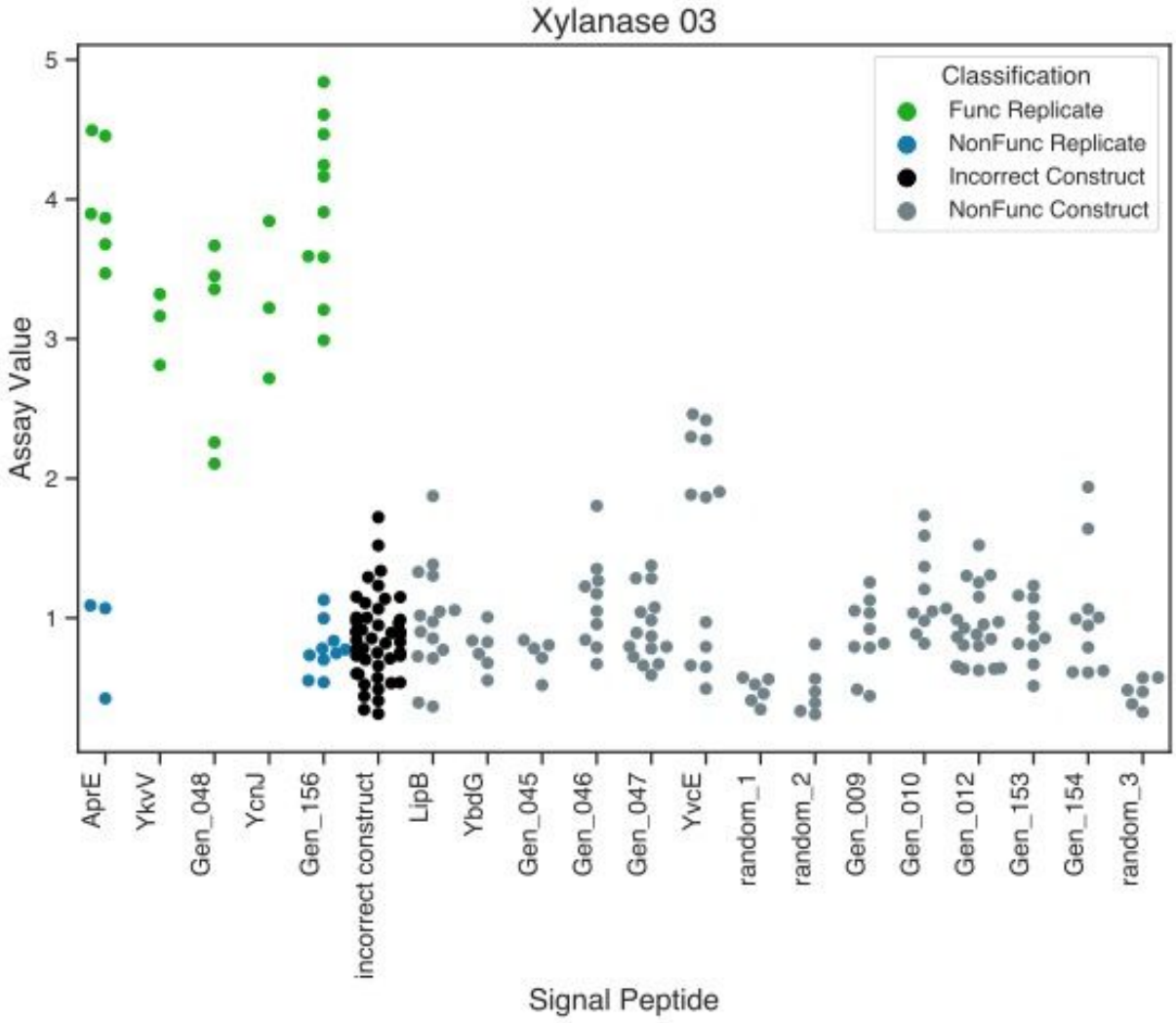
Assay information is provided in Supplemental Table 2.



Shuffled_ID: seq36
Original_ID: Arnold_034
Plot_ID: Protease 03

Construct Functionality Classification Summary
No constructs with SPs generated for random inputs tested.
No constructs with natural SPs tested.
0.18 : 2 out of 11 constructs with generated SPs functional

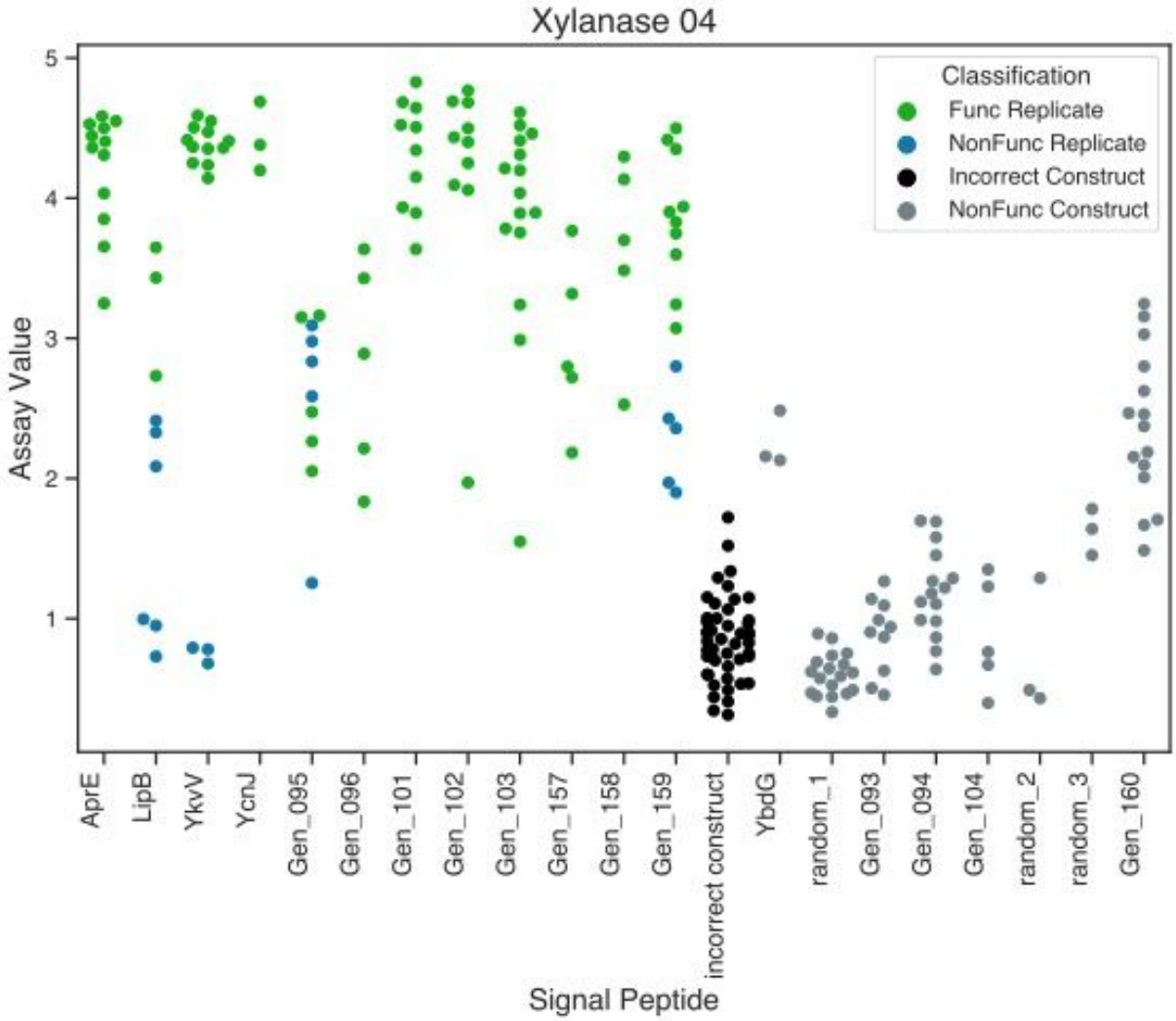
Assay information is provided in Supplemental Table 2.



Shuffled_ID: seq3
Original_ID: Arnold_039
Plot_ID: Xylanase 03

Construct Functionality Classification Summary
0.00 : 0 out of 3 constructs with SPs generated for random inputs functional
0.50 : 3 out of 6 constructs with natural SPs functional
0.20 : 2 out of 10 constructs with generated SPs functional

Assay information is provided in Supplemental Table 2.



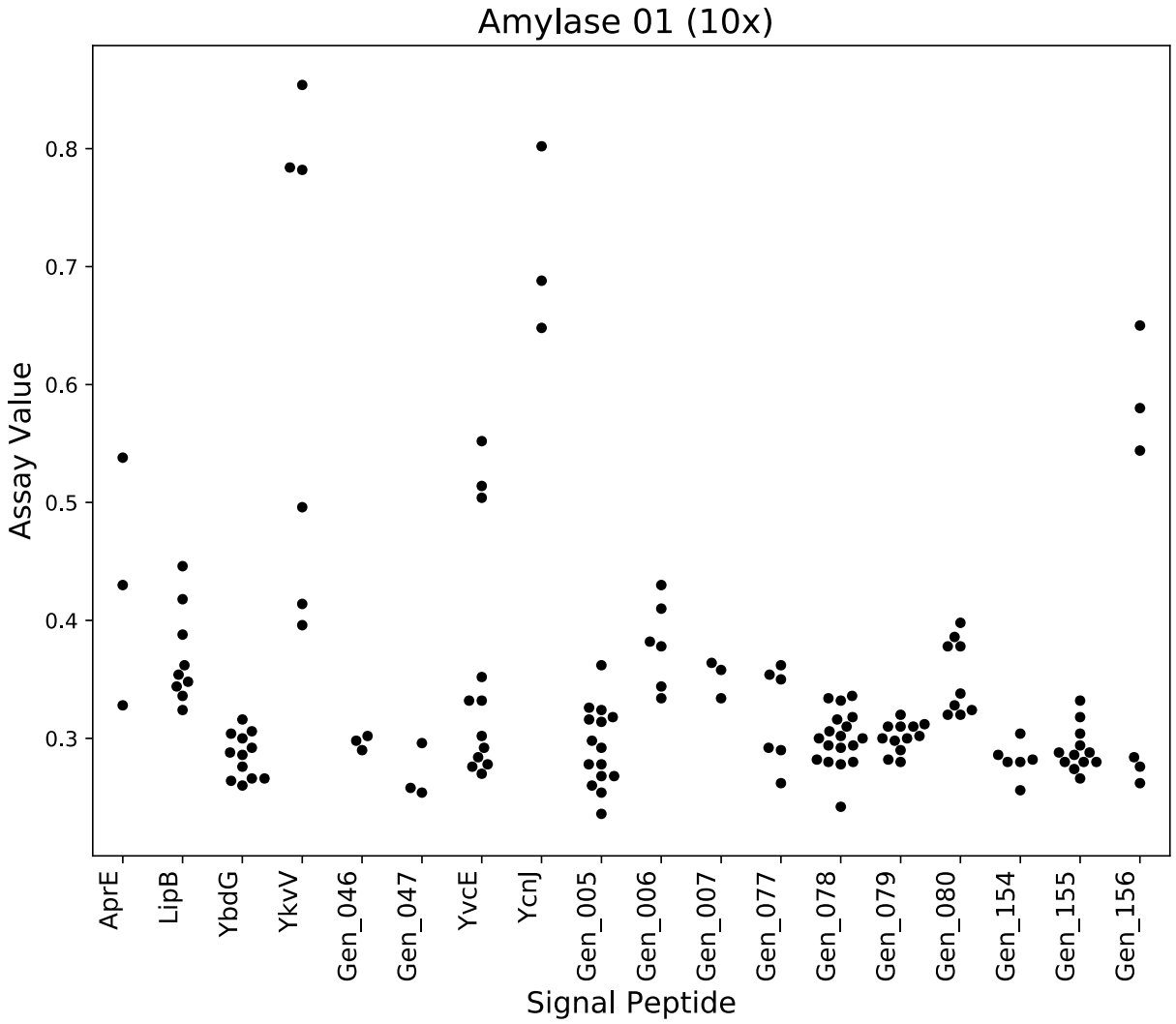
Shuffled_ID: seq26
Original_ID: Arnold_040
Plot_ID: Xylanase 04

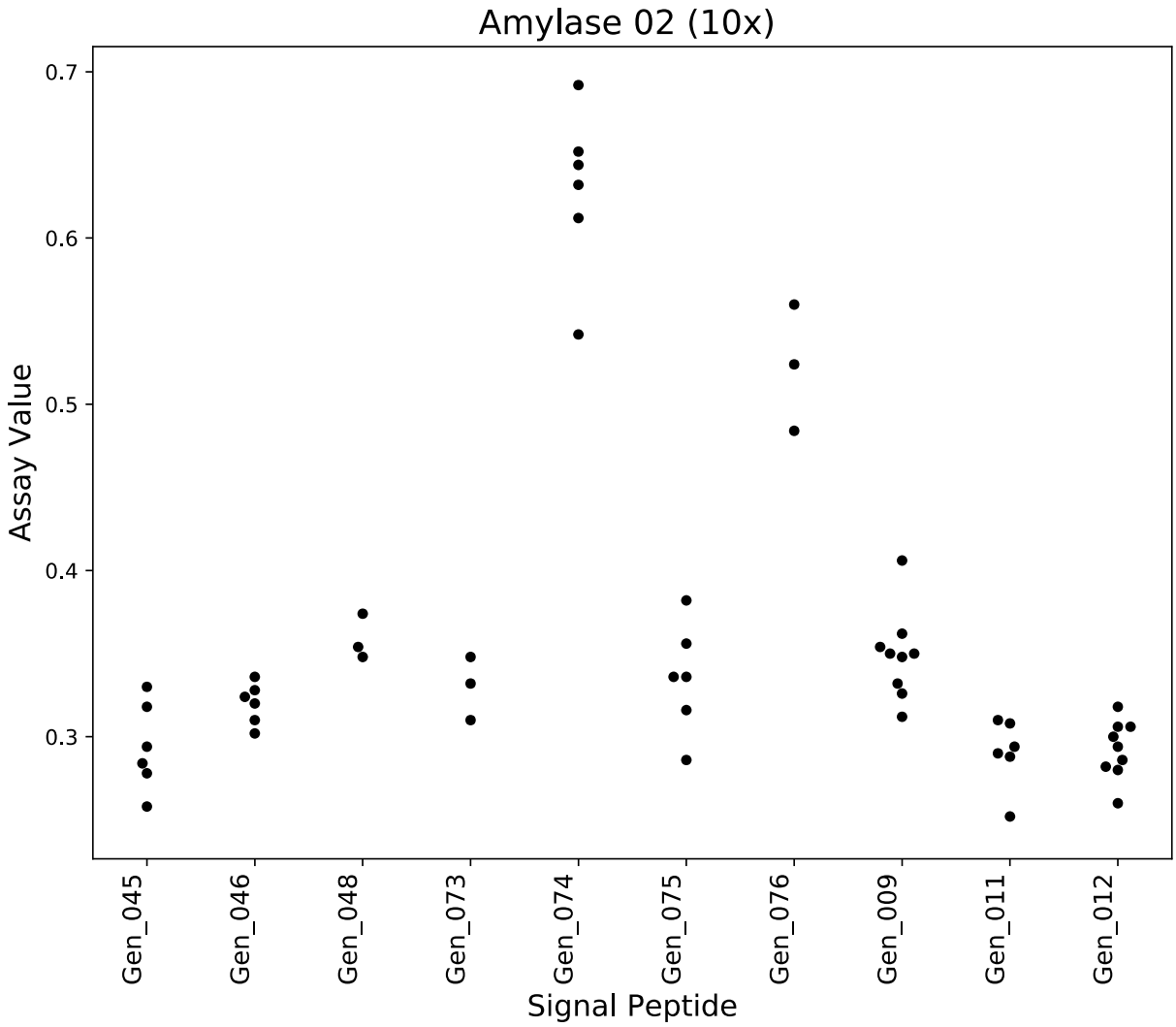
Construct Functionality Classification Summary
0.00 : 0 out of 3 constructs with SPs generated for random inputs functional
0.80 : 4 out of 5 constructs with natural SPs functional
0.67 : 8 out of 12 constructs with generated SPs functional

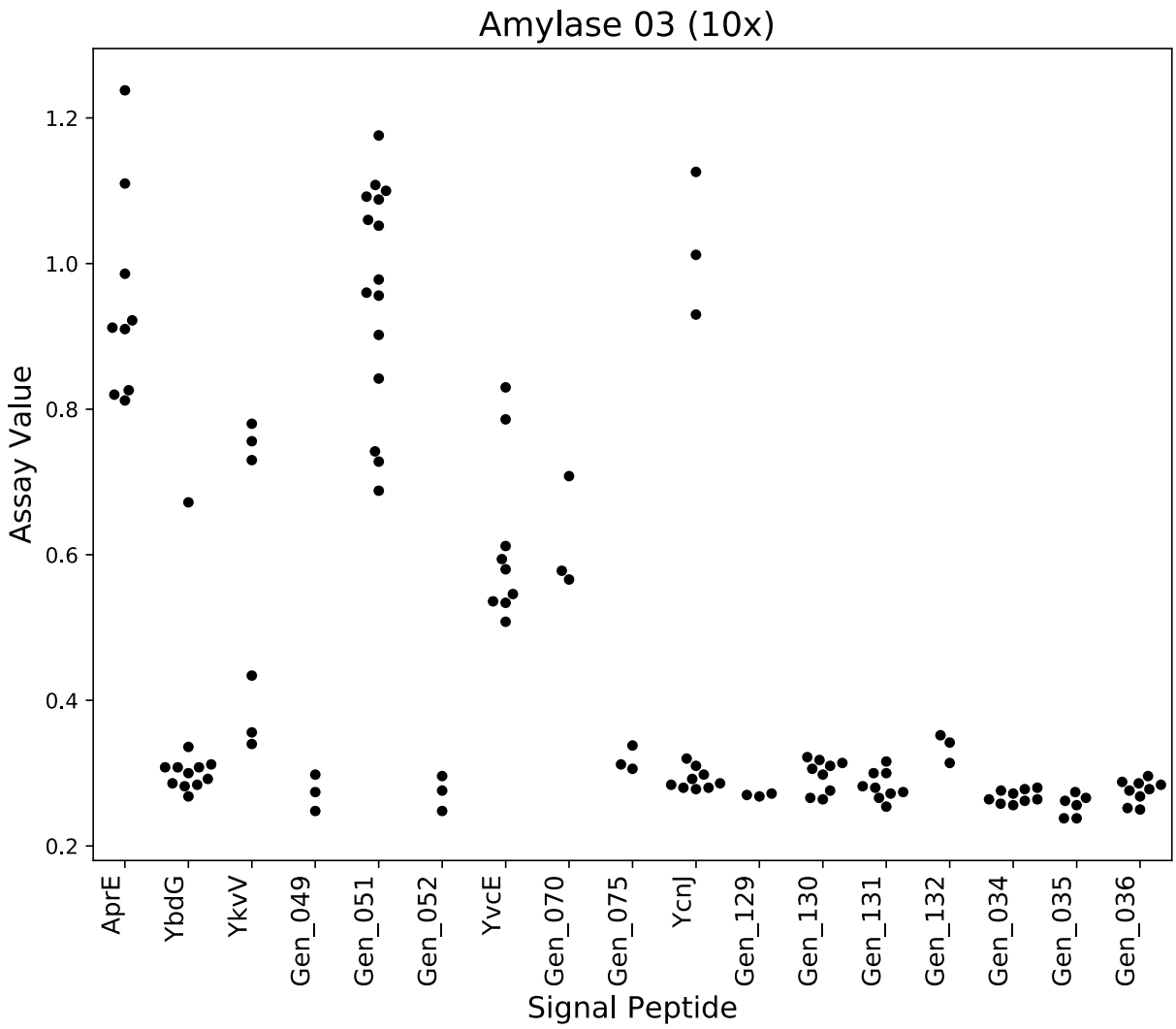
Assay information is provided in Supplemental Table 2.

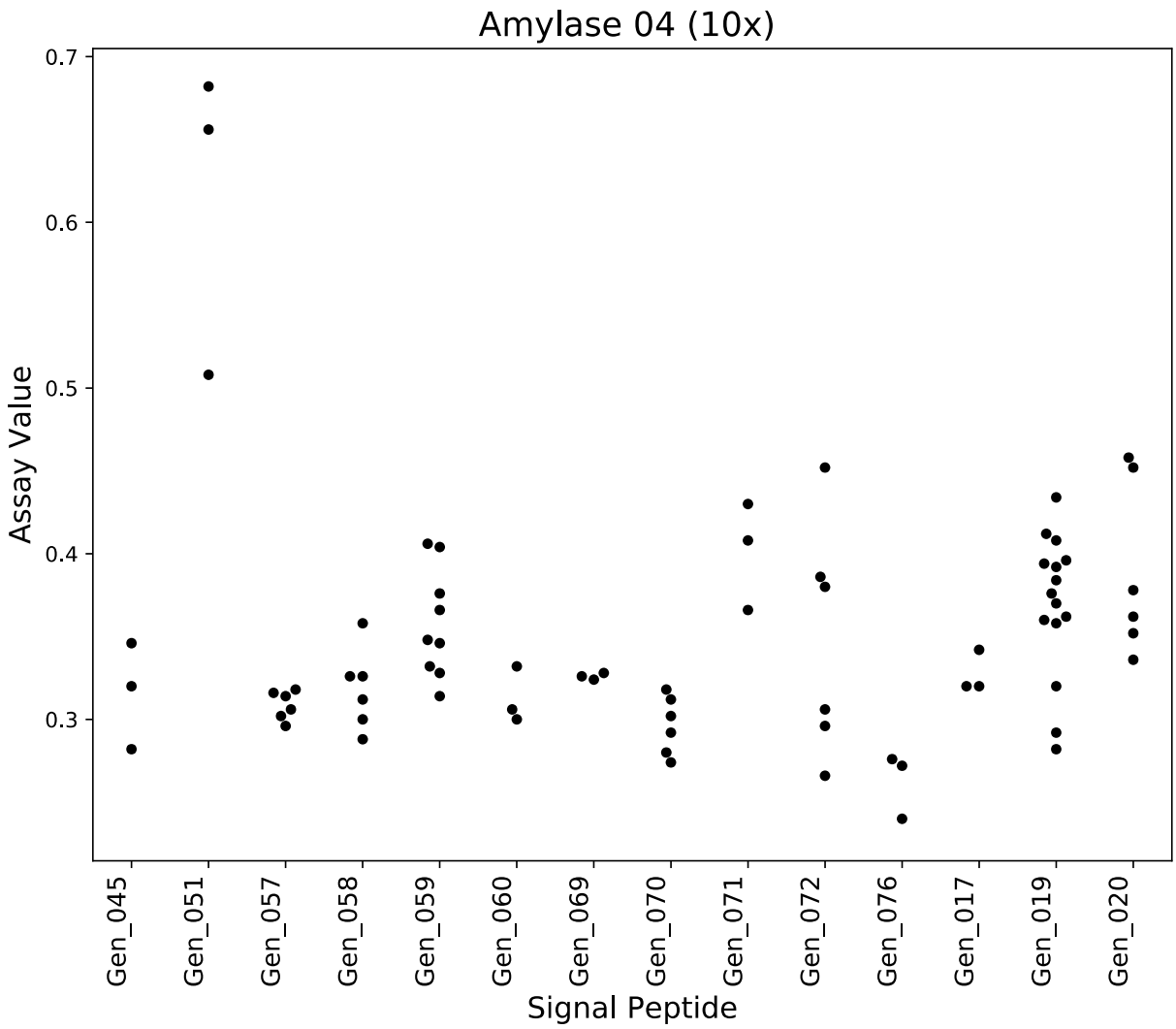
Supp Section III: Activity assays at higher dilution

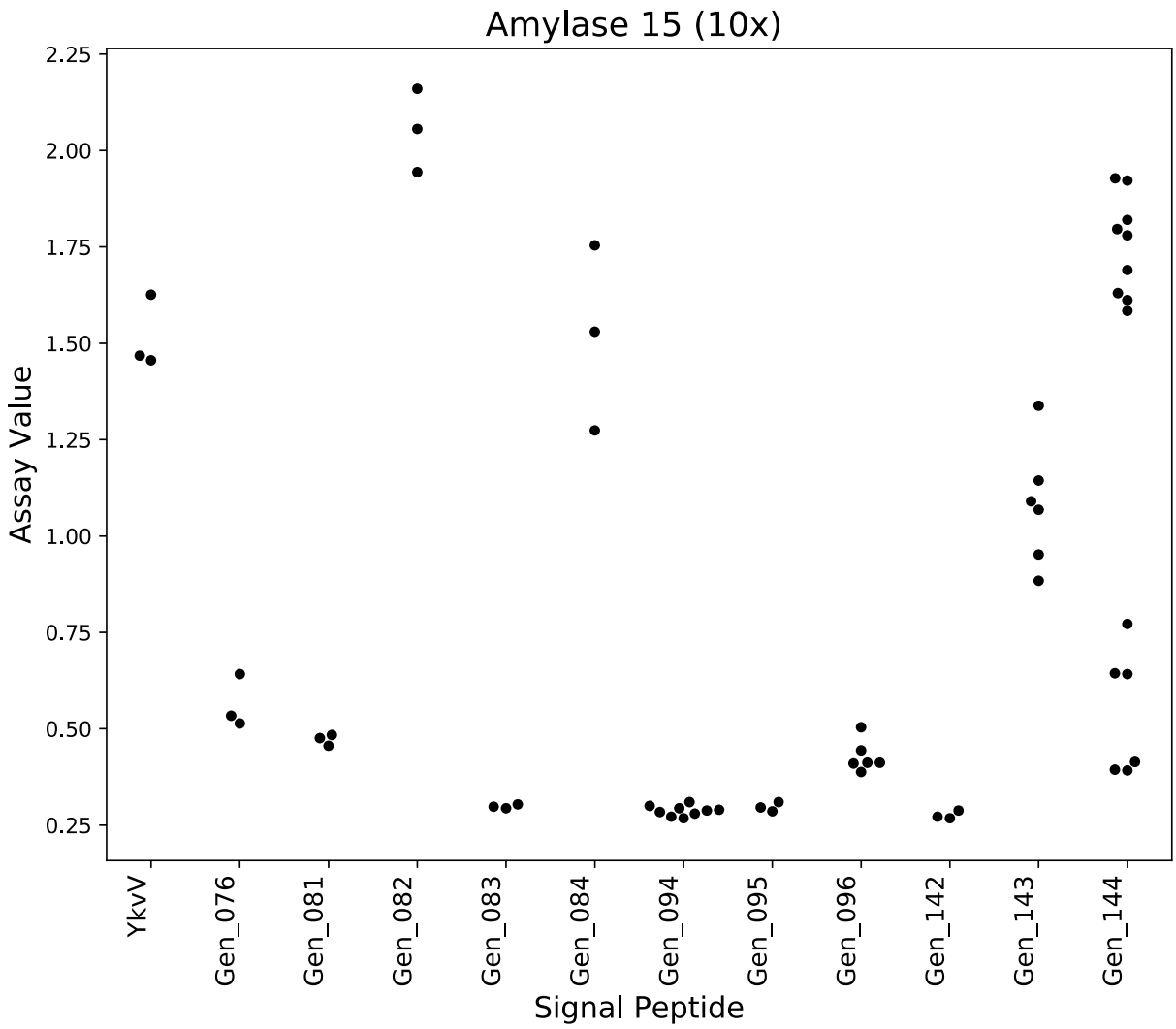
While Supplemental Section 2 shows activity data for identifying functional constructs, we found that some amylase and xylanase constructs were outside of the linear range (activity was too high) to make accurate comparisons between the high functioning constructs. To address this, we performed the assays at higher dilution (10x for amylase, 50x for xylanase) to obtain data in the linear range of the assay.

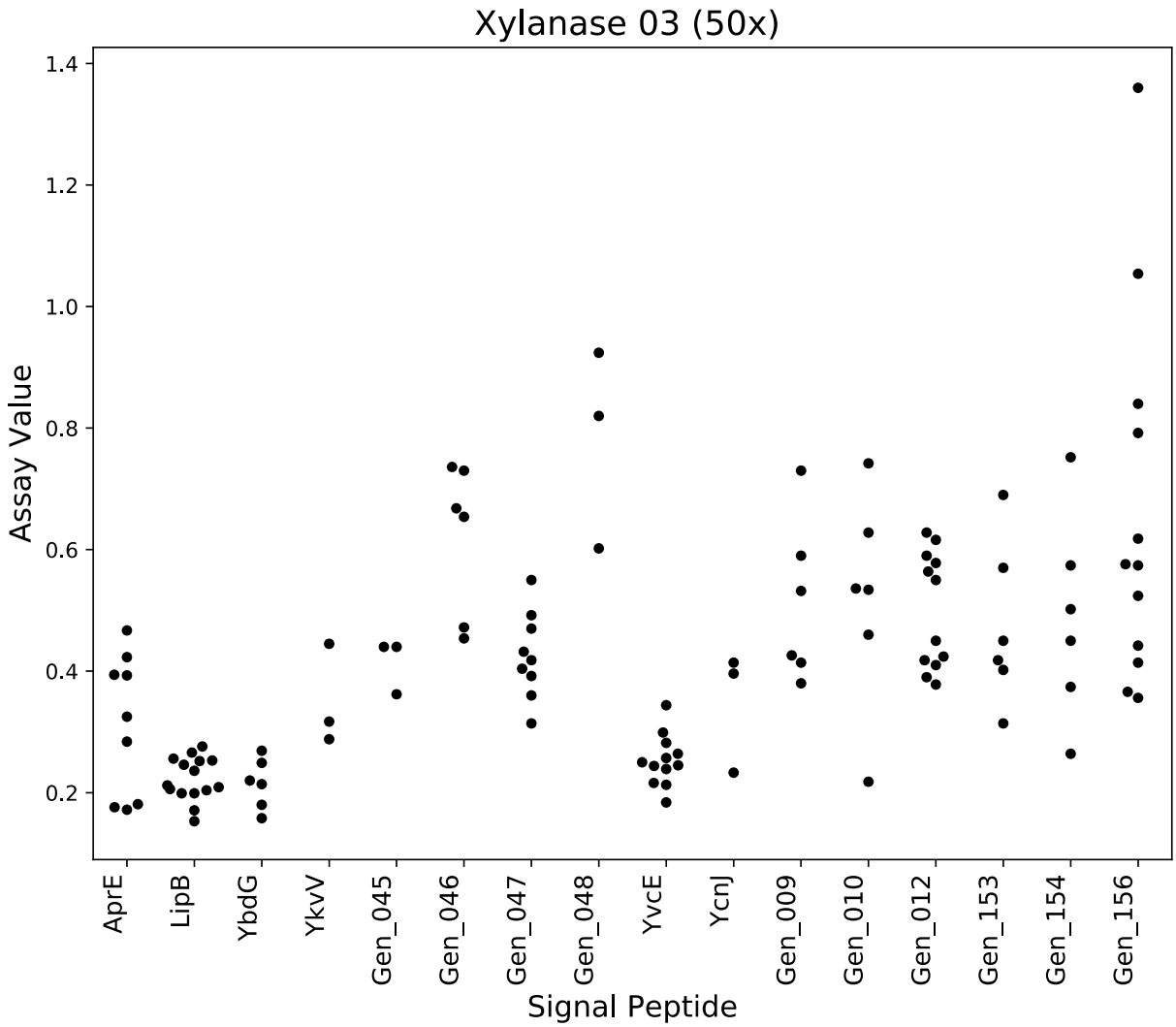


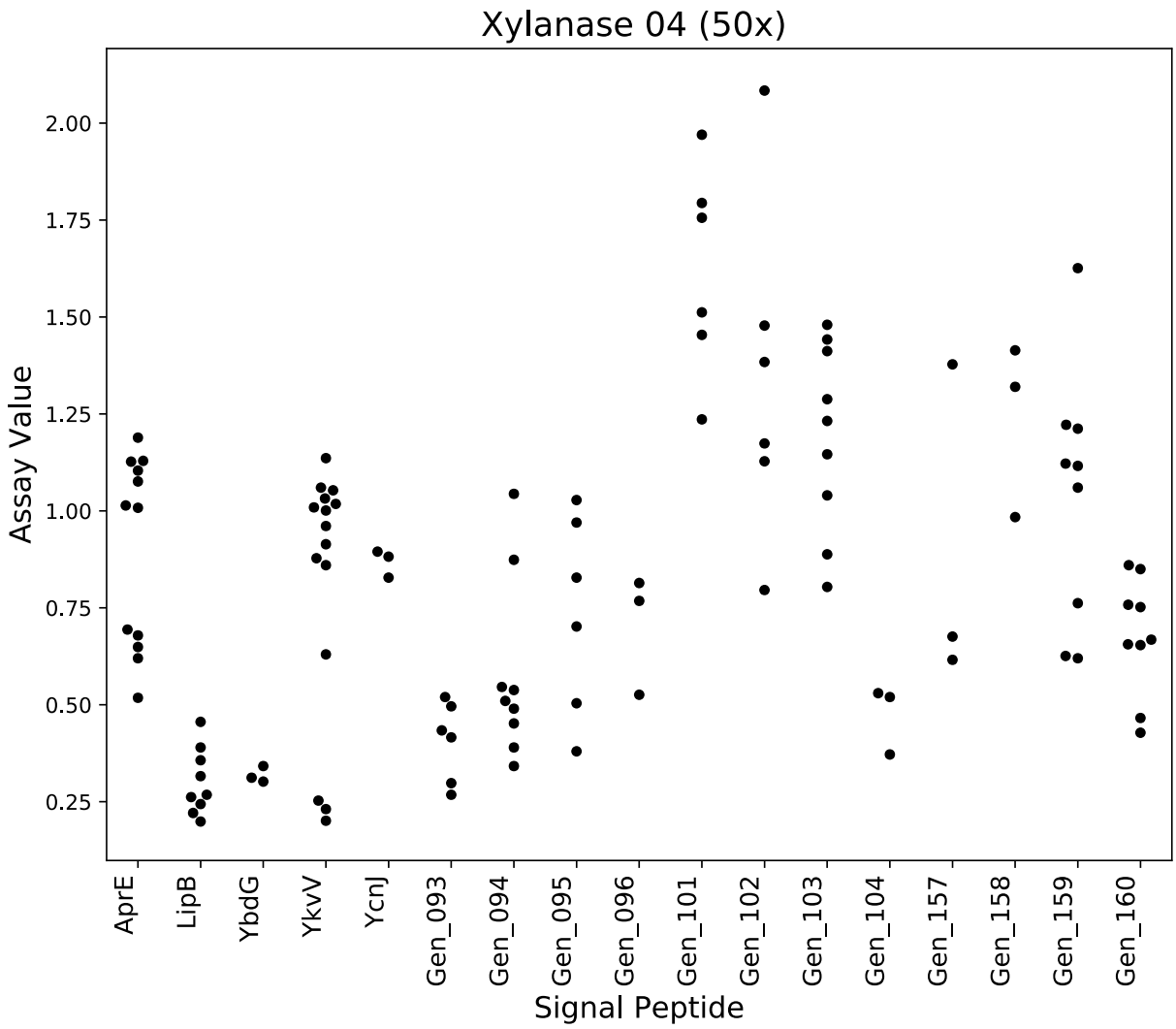

Amylase 02 (10x)
Assay Value
0.7
0.6
0.5
0.4
0.3
Gen_045
Gen_046
Gen_048
Gen_073
Gen_074
Gen_075
Gen_076
Gen_009
Gen_011
Gen_012
Signal Peptide


Amylase 03 (10x)
Assay Value
1.2
1.0
0.8
0.6
0.4
0.2
AprE
YbdG
YkvV
Gen_049
Gen_051
Gen_052
YvcE
Gen_070
Gen_075
YcnJ
Gen_129
Gen_130
Gen_131
Gen_132
Gen_034
Gen_035
Gen_036
Signal Peptide


Amylase 04 (10x)
Assay Value
0.7
0.6
0.5
0.4
0.3
Gen_045
Gen_051
Gen_057
Gen_058
Gen_059
Gen_060
Gen_069
Gen_070
Gen_071
Gen_072
Gen_076
Gen_017
Gen_019
Gen_020
Signal Peptide


Amylase 15 (10x)
Assay Value
2.25
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
YkvV
Gen_076
Gen_081
Gen_082
Gen_083
Gen_084
Gen_094
Gen_095
Gen_096
Gen_142
Gen_143
Gen_144
Signal Peptide


Xylanase 03 (50x)
Assay Value
1.4
1.2
1.0
0.8
0.6
0.4
0.2
AprE
LipB
YbdG
YkvV
Gen_045
Gen_046
Gen_047
Gen_048
YvcE
YcnJ
Gen_009
Gen_010
Gen_012
Gen_153
Gen_154
Gen_156
Signal Peptide


Xylanase 04 (50x)
Assay Value
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
AprE
LipB
YbdG
YkvV
YcnJ
Gen_093
Gen_094
Gen_095
Gen_096
Gen_101
Gen_102
Gen_103
Gen_104
Gen_157
Gen_158
Gen_159
Gen_160
Signal Peptide

Supp Section IV: Characteristics of functional vs nonfunctional generated SPs

We attempted to identify general characteristics of functional versus nonfunctional signal peptides that may help in discerning whether a signal peptide is functional or not. To do so, we separated the 27 signal peptides that are only present in functional constructs, and 30 signal peptides that are only present in nonfunctional constructs, and attempted to identify a distinguishing property. No property was identified for distinguishing between functional and nonfunctional constructs (p-value < 0.05) and SignalP 5.0 was also not able to discern between functional and nonfunctional generated SPs (ROC=0.59).

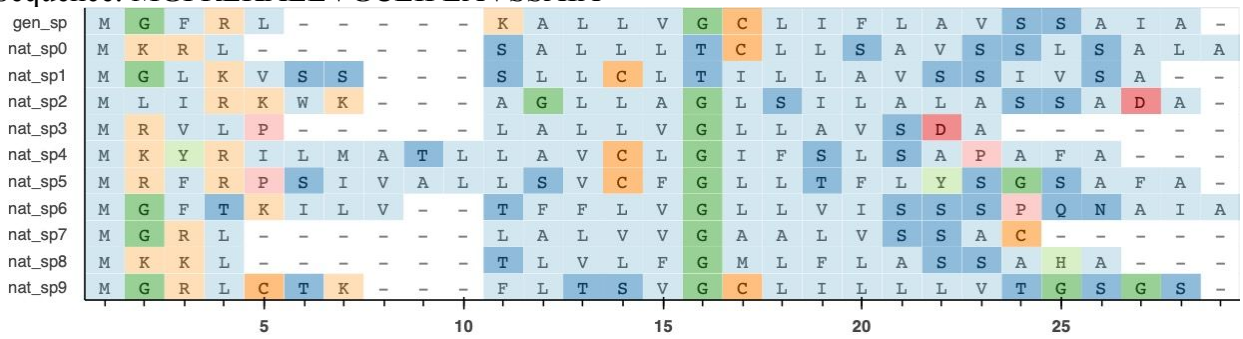
Protein Property	**p-value**
Length	0.9229
Hydrophobicity Index	0.3870
Hydrophilicity Index	0.2126
Molecular Weight (MW)	0.7963
Average MW	0.1170
Aromaticity	0.9935
Instability Index	0.9748
Isoelectric Point	0.3243

Supp Section V: All MSAs for functional SPs

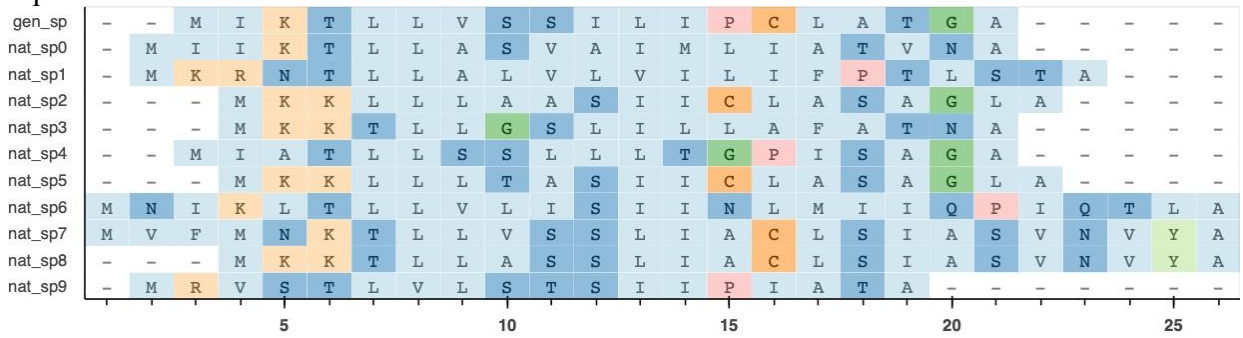
MSAs for the top 10 closest matching natural Signal Peptides (by sequence identity) to each of the 43 generated SPs that were functional in any construct are provided. Alignments are sorted by closest sequence identity to the generated natural Signal Peptide (shown first).

All generated Signal Peptides can be found in Supplemental File 1.

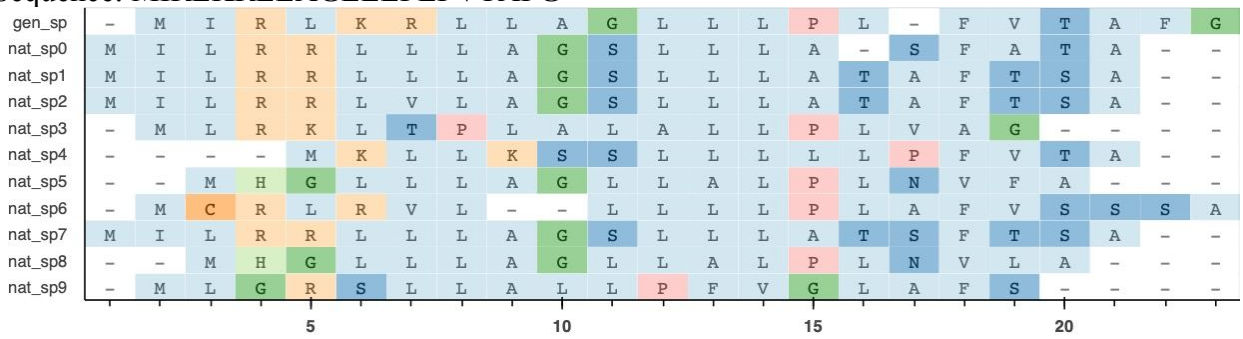
Sequence: MGFRLKALLVGCLIFLAVSSAIA



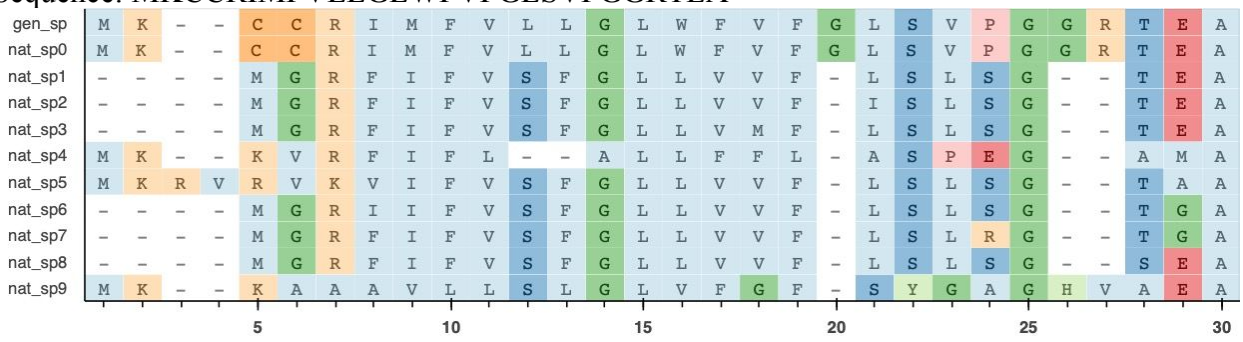
Sequence: MIKTLLVSSILIPCLATGA



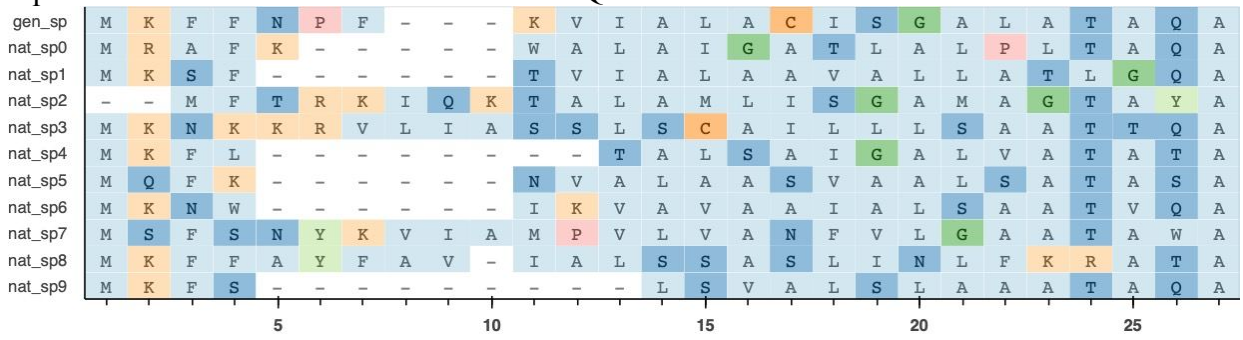
Sequence: MIRLKRLLAGLLLPLFVTAFG



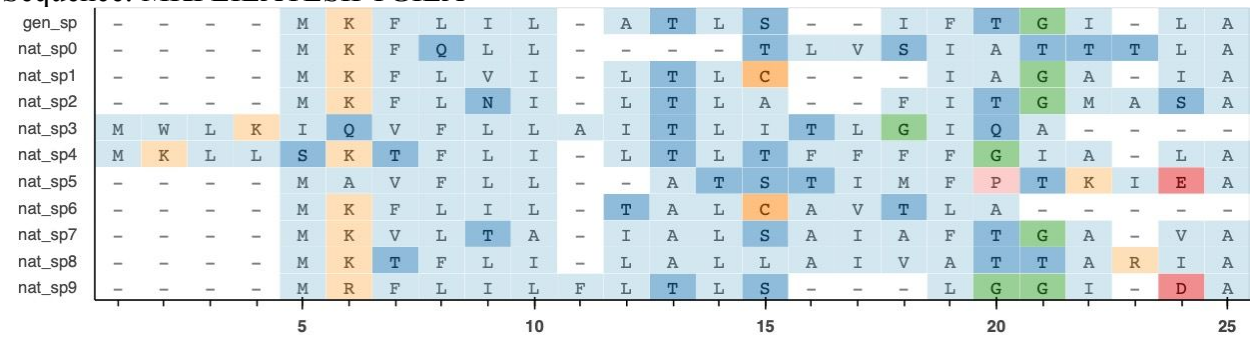
Sequence: MKCCRIMFVLLGLWFVFGLSVPGGRTEA



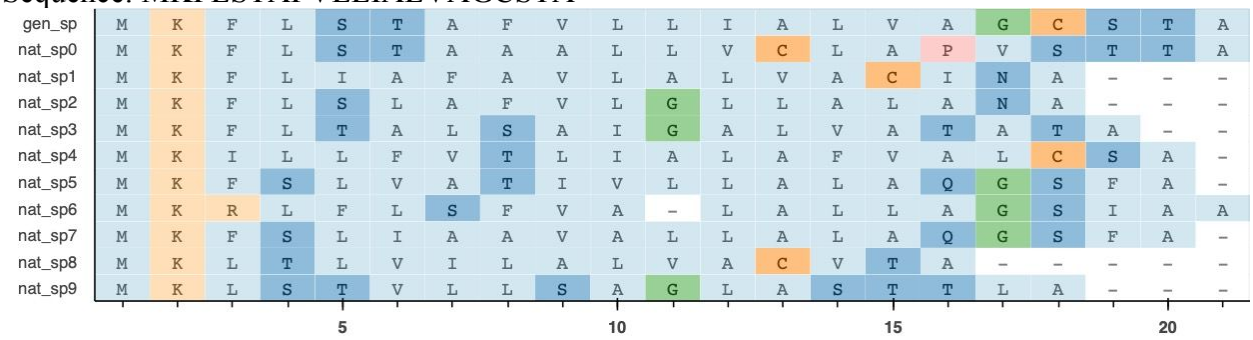
Sequence: MKFFNPFKVIALACISGALATAQA



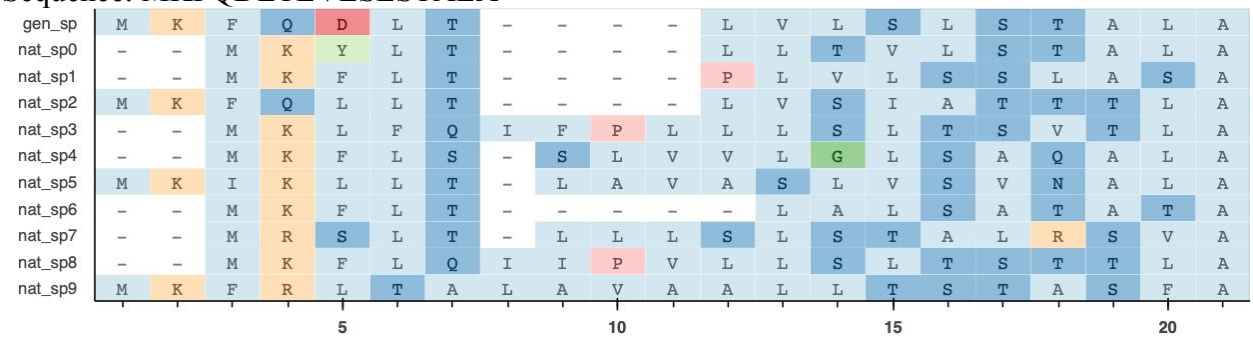
Sequence: MKFLILATLSIFTGILA



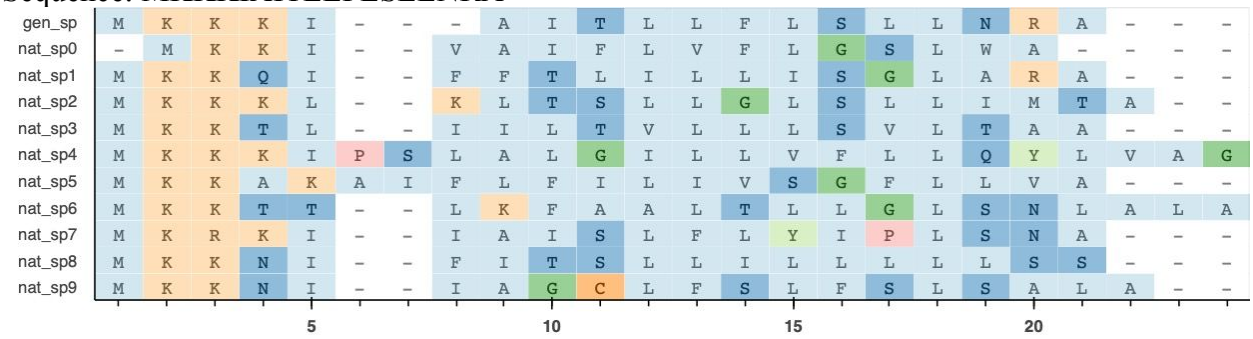
Sequence: MKFLSTAFVLLIALVAGCSTA



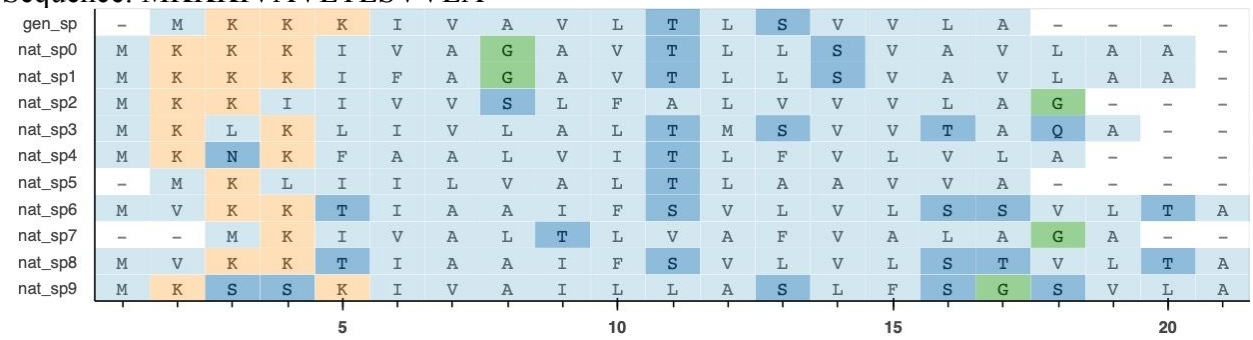
Sequence: MKFQDLTLVLSLSTALA



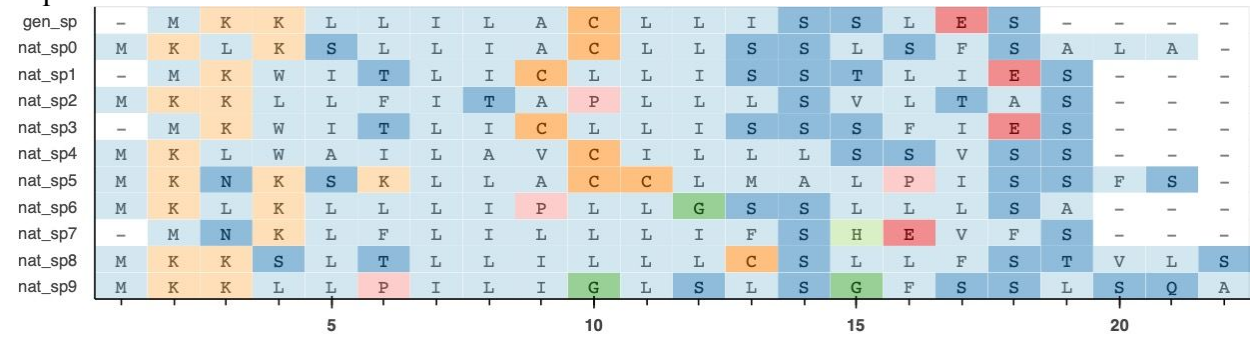
Sequence: MKKKIAITLLFLSLLNRA



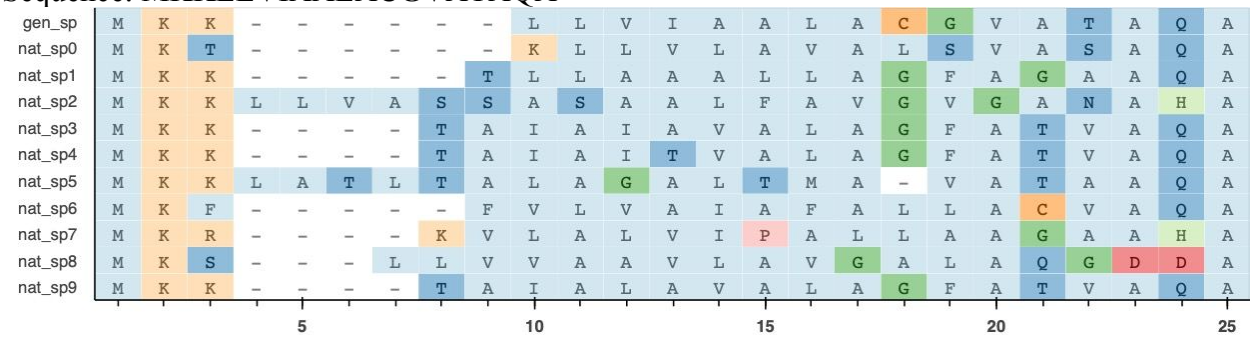
Sequence: MKKKIVAVLTLSVVLA



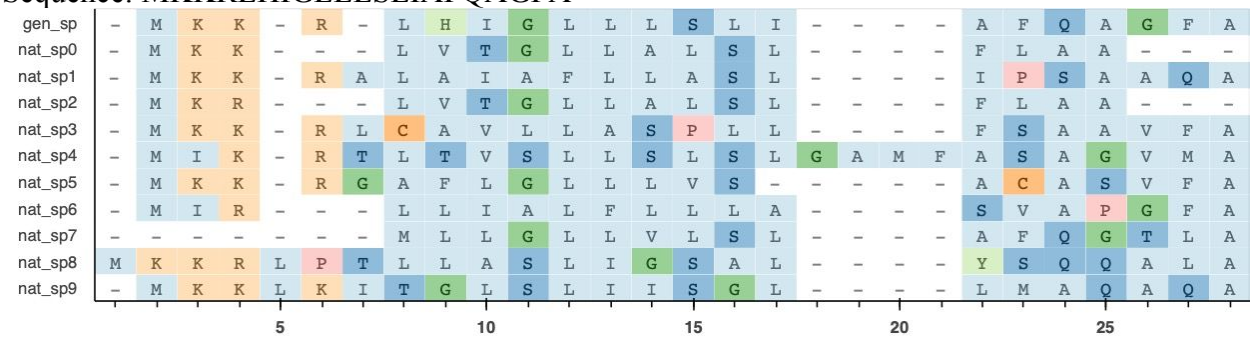
Sequence: MKKLLILACLLISSLES



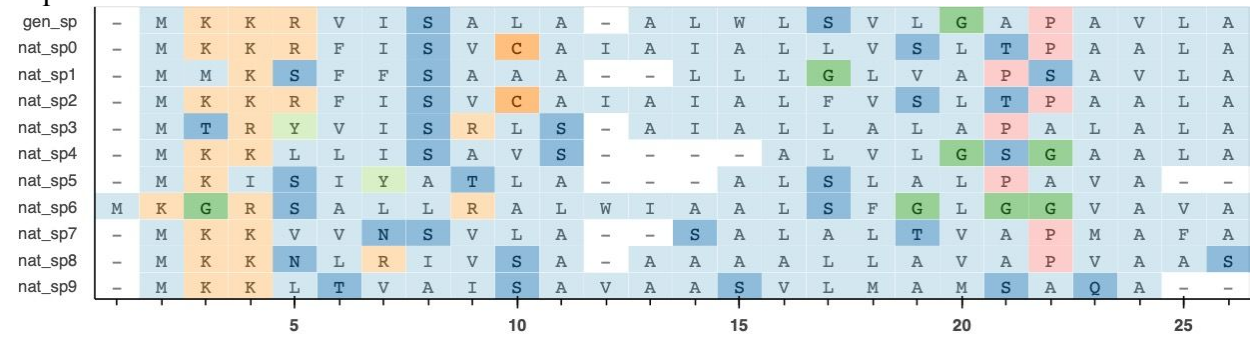
Sequence: MKKLLVIAALACGVATAQA



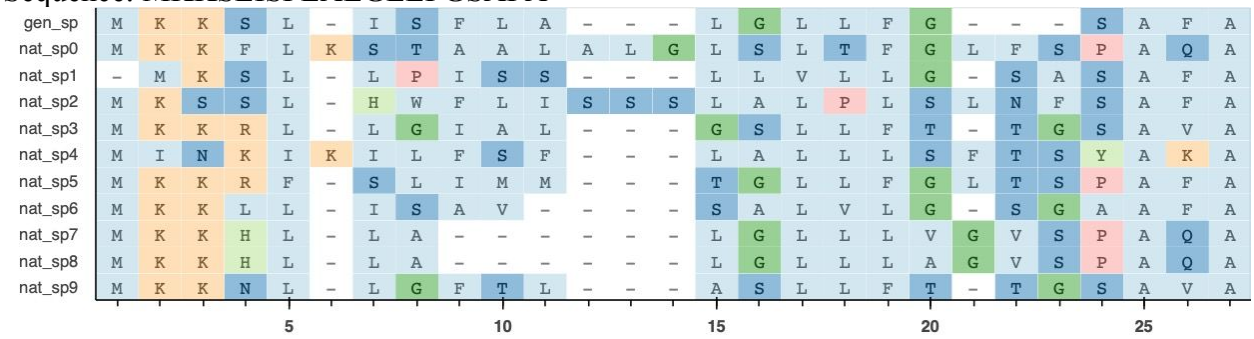
Sequence: MKKRLHIGLLLSLIAFQAGFA



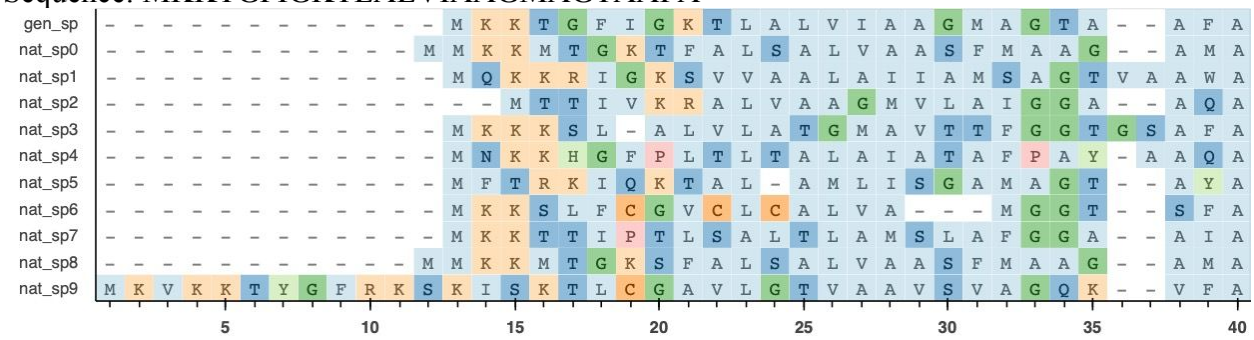
Sequence: MKKRVISALAALWLSVLGAPAVLA



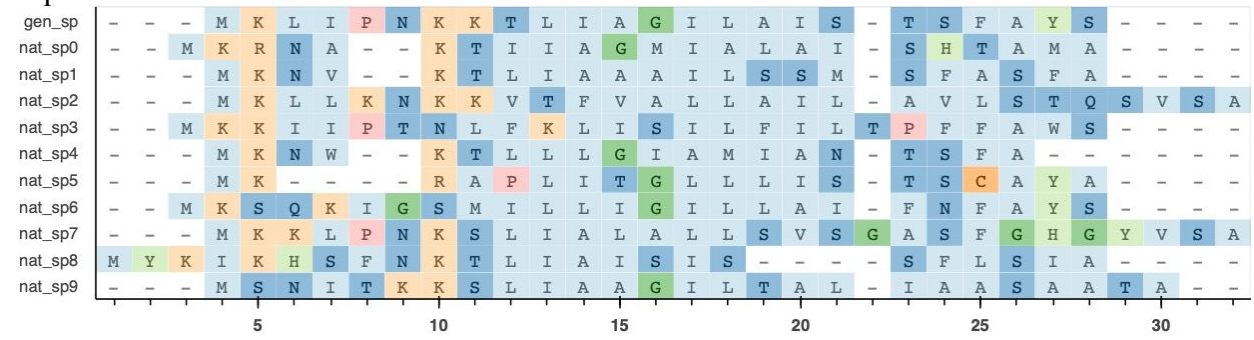
Sequence: MKKSLISFLALGLLFGSAFA



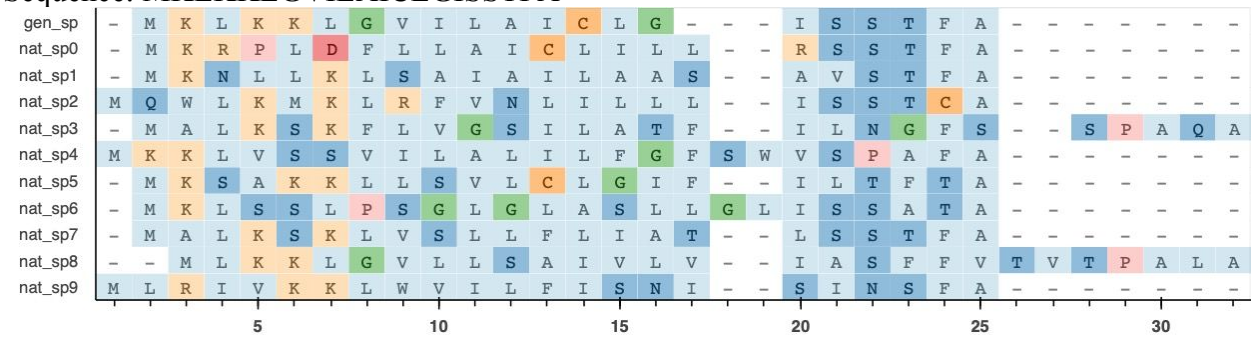
Sequence: MKKTGFIGKTLALVIAAGMAGTAAFA



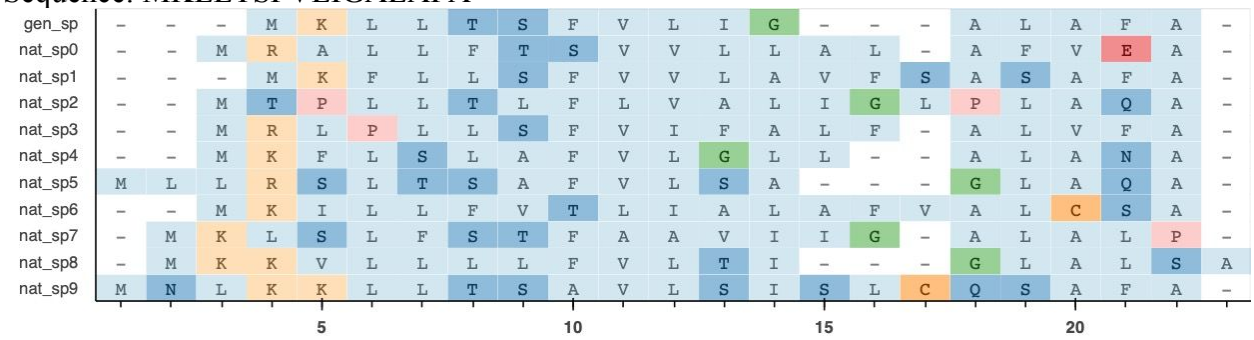
Sequence: MKLIPNKKTLIAGILAISTSFAYS



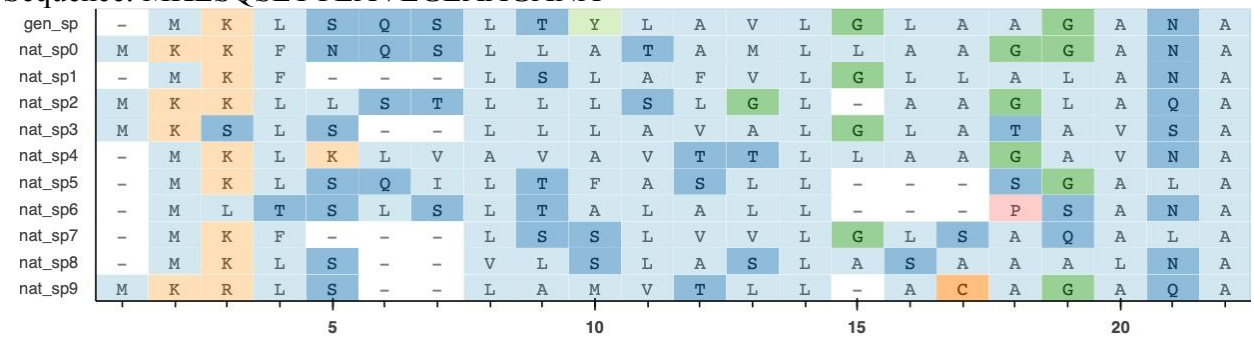
Sequence: MKLKKLGVILAICLGISSTFA



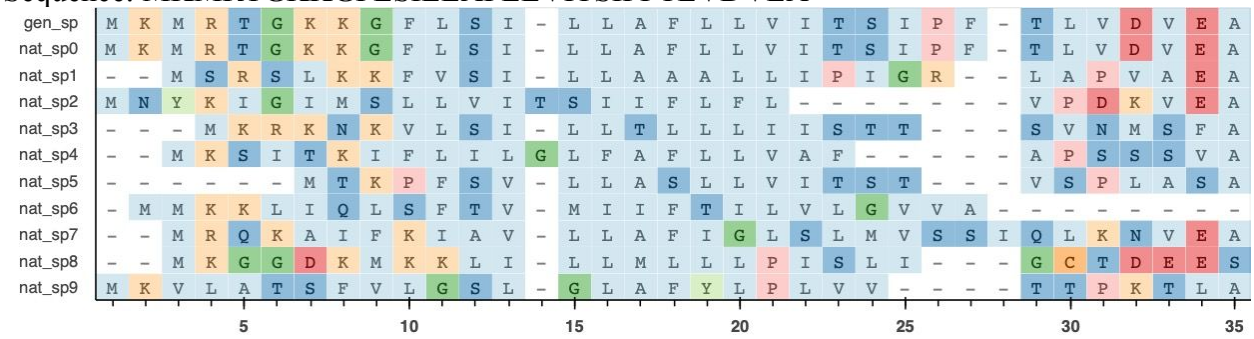
Sequence: MKLLTSFVLIGALAFA



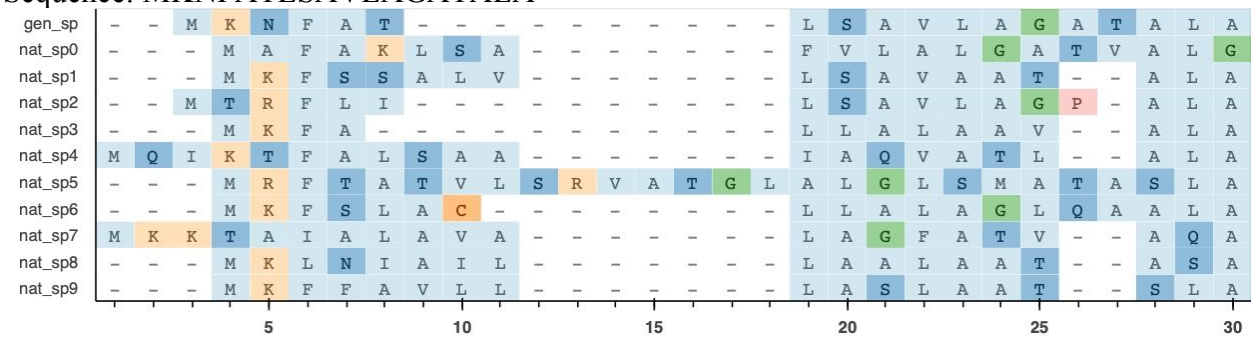
Sequence: MKLSQSLTYLAVLGLAAGANA



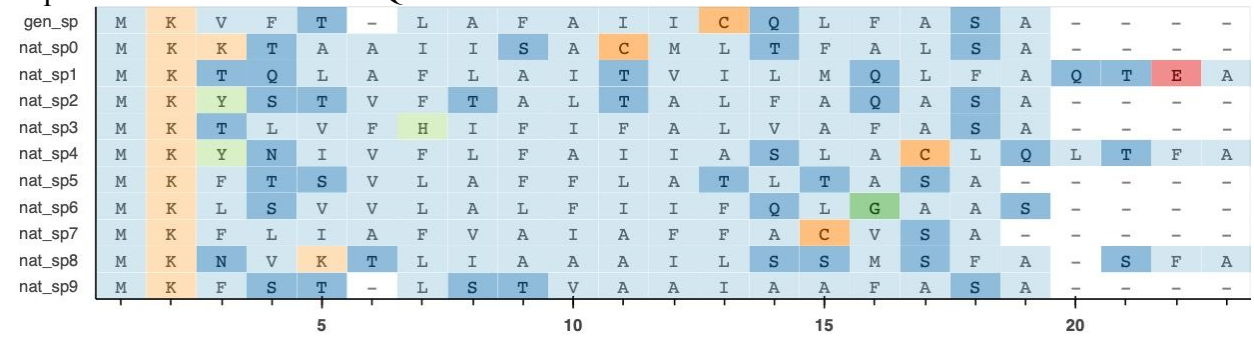
Sequence: MKMRTGKKGFLSILLAFLLVITSIPFTLVDVEA



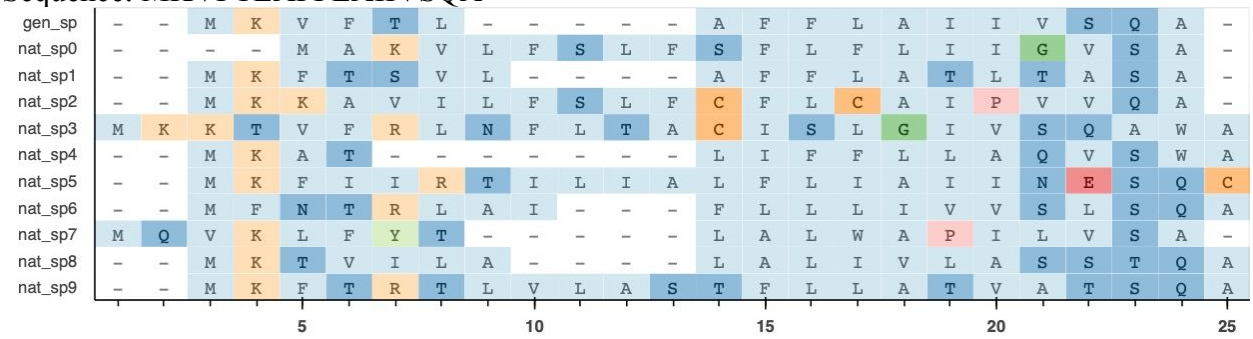
Sequence: MKNFATLSAVLAGATALA



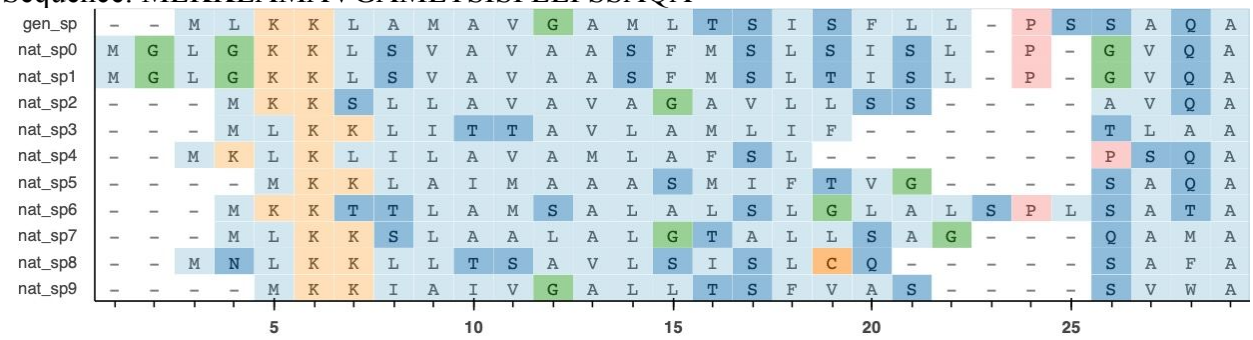
Sequence: MKVFTLAFAIICQLFASA



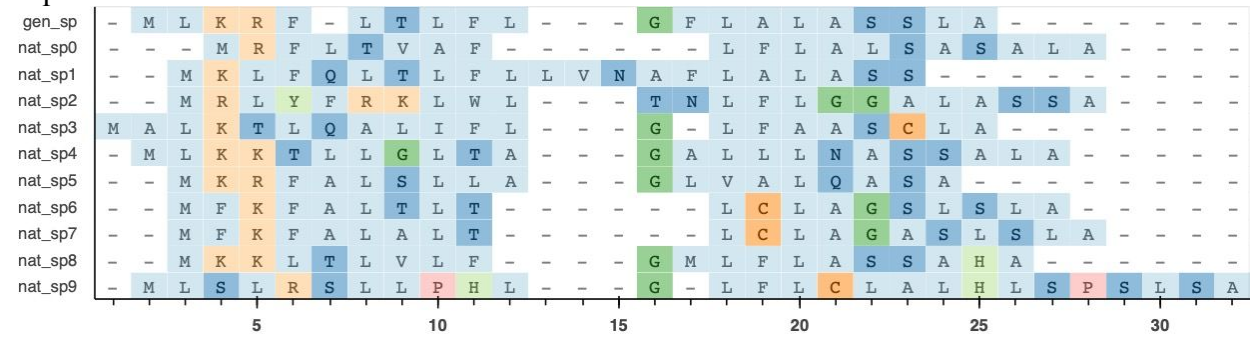
Sequence: MKVFTLAFFLAIIVSQA



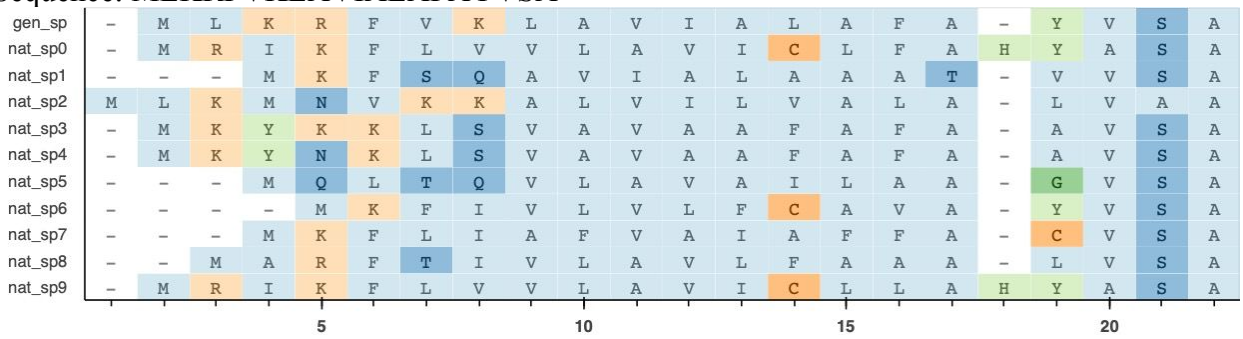
Sequence: MLKKLAMAVGAMLTSISFLLPSSAQA



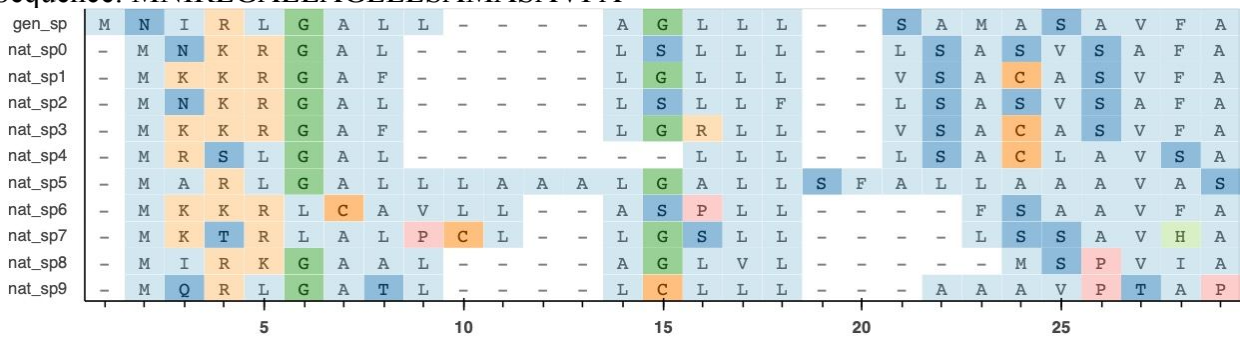
Sequence: MLKRFLTLFLGFLALASSLA



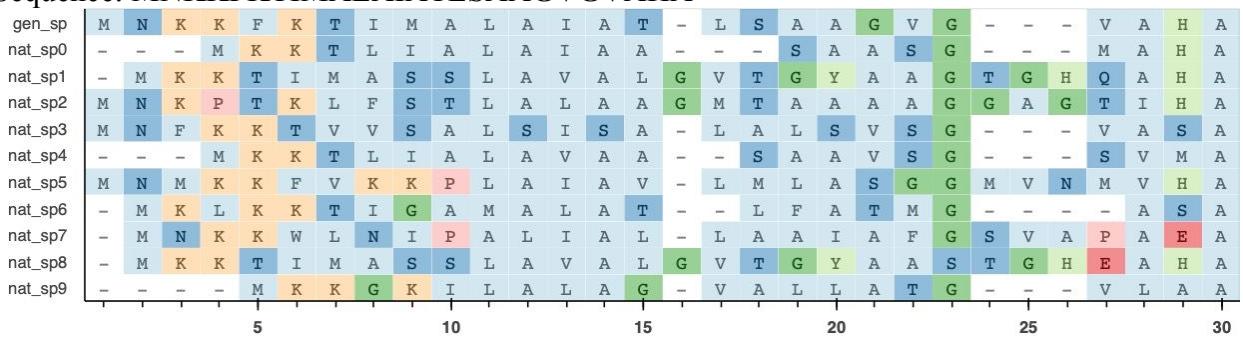
Sequence: MLKRFVKLAVIALAFAYVSA



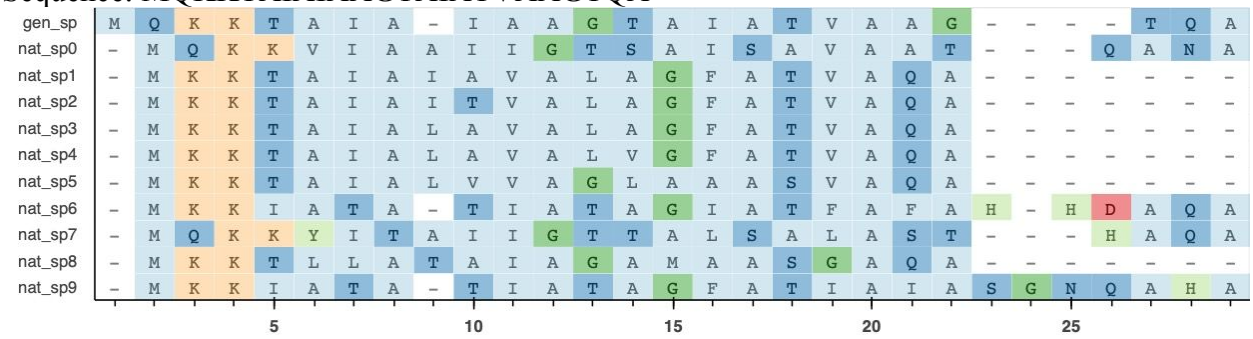
Sequence: MNIRLGALLAGLLLSAMASAVFA



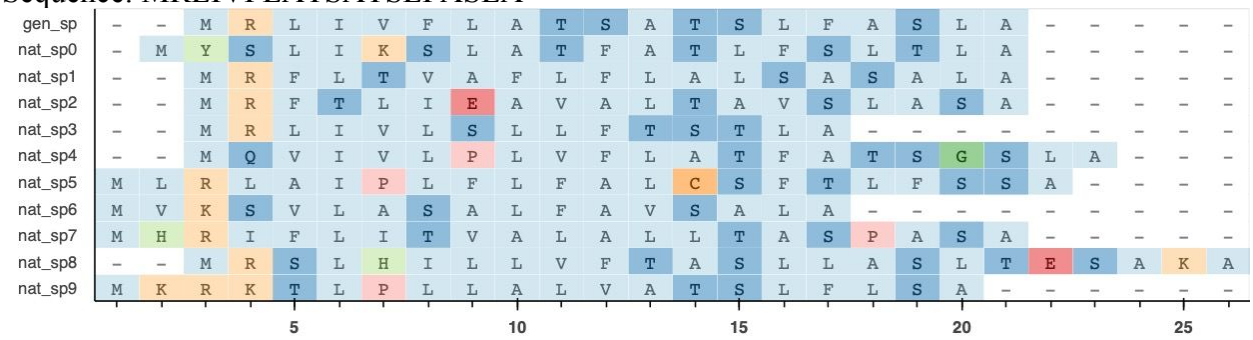
Sequence: MNKKFKTIMALAIATLSAAGVGVAHA



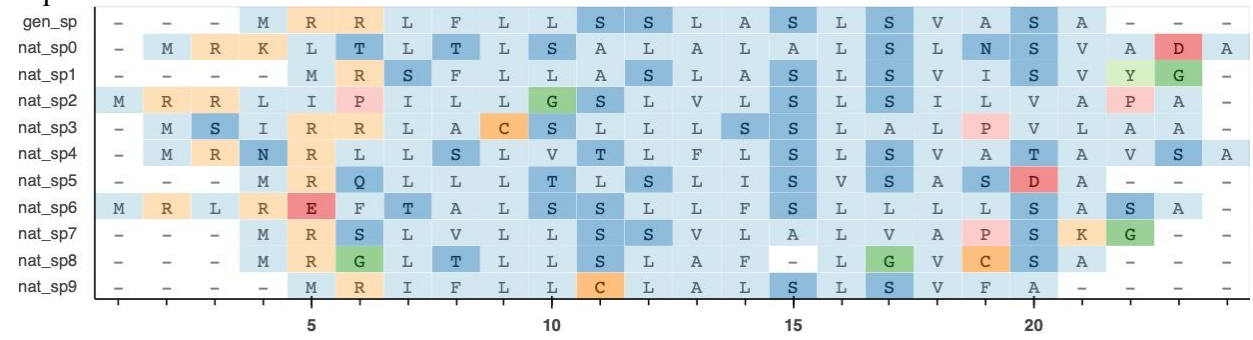
Sequence: MQKKTAIAIAAGTAIATVAAGTQA



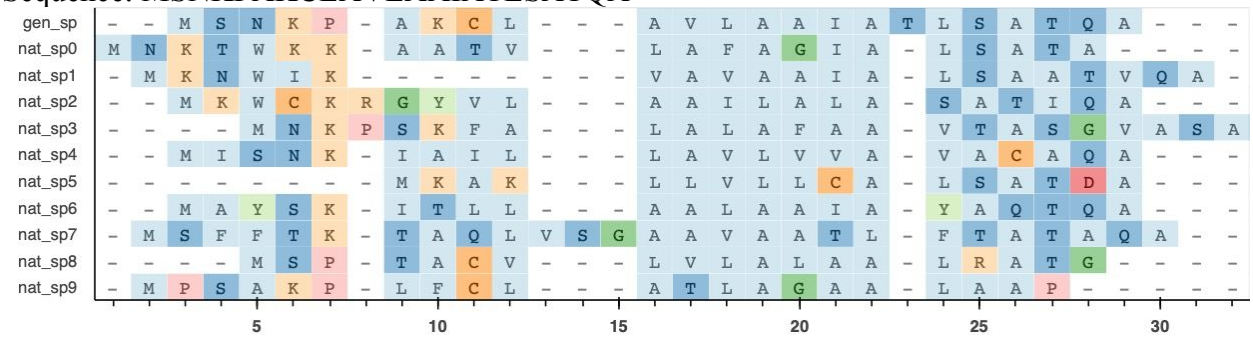
Sequence: MRLIVFLATSATSLFASLA



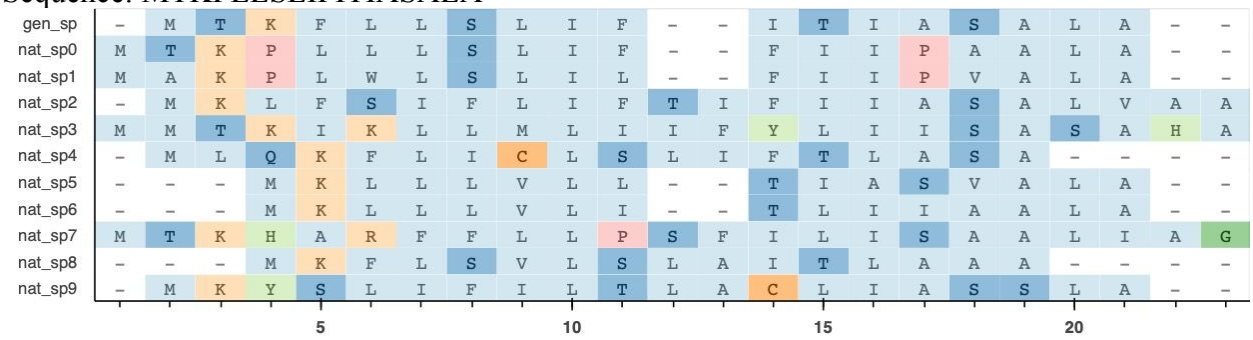
Sequence: MRRLFLLSSLASLSVASA



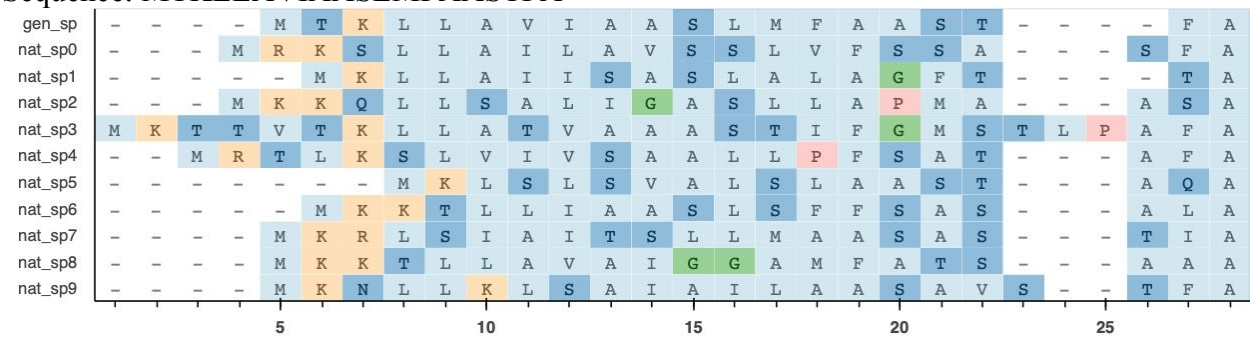
Sequence: MSNKPAKCLAVLAAIATLSATQA



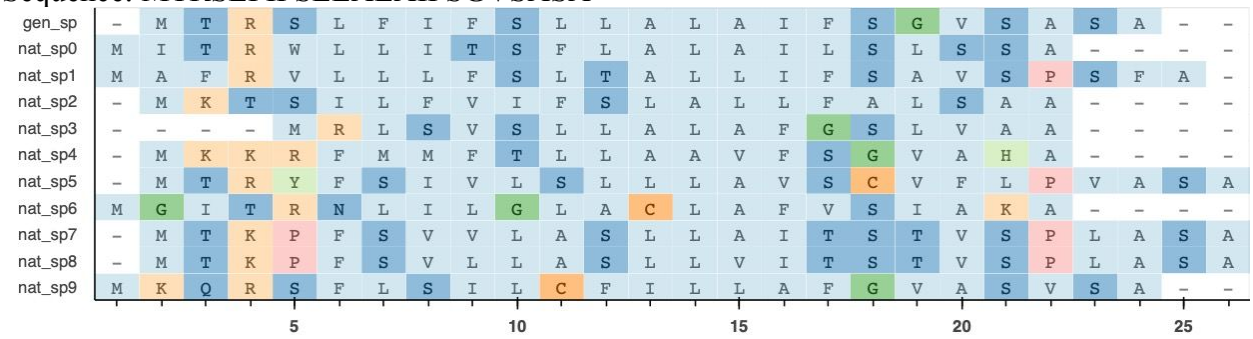
Sequence: MTKFLLSLIFITIASALA



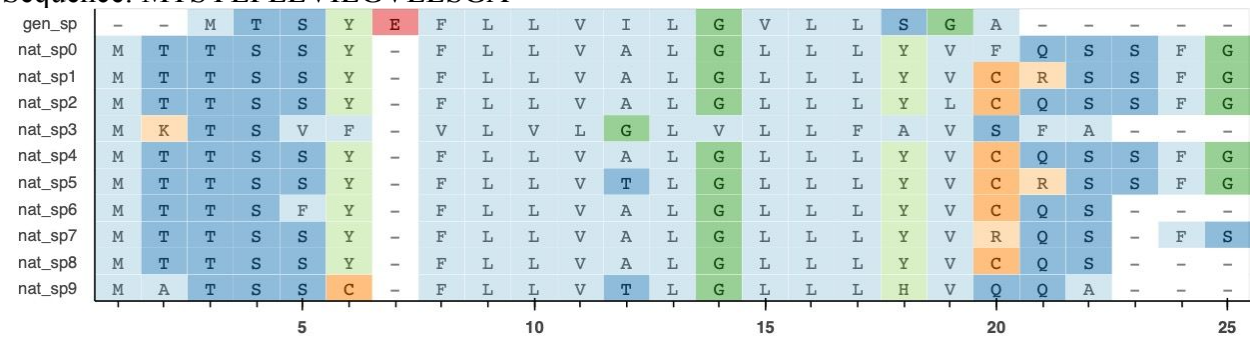
Sequence: MTKLLAVIAASLMFAASTFA



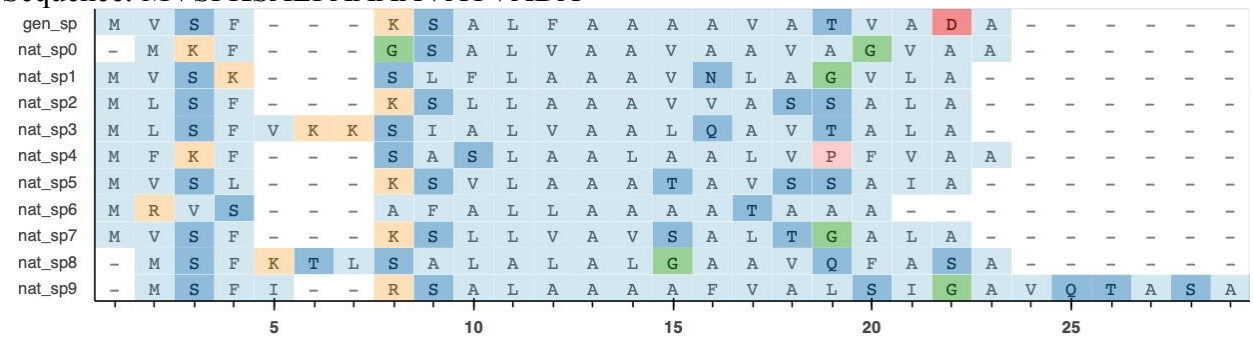
Sequence: MTRSLFIFSLLALAIFSGVSASA



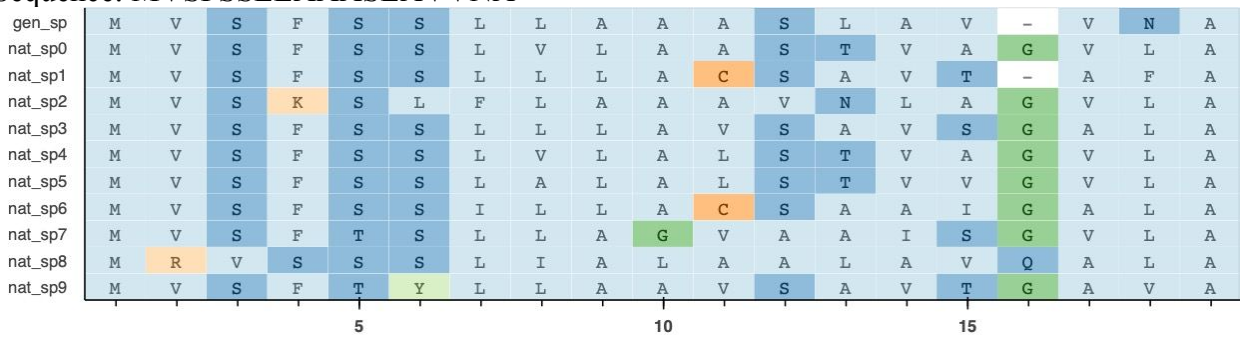
Sequence: MTSYEFLLVILGVLLSGA



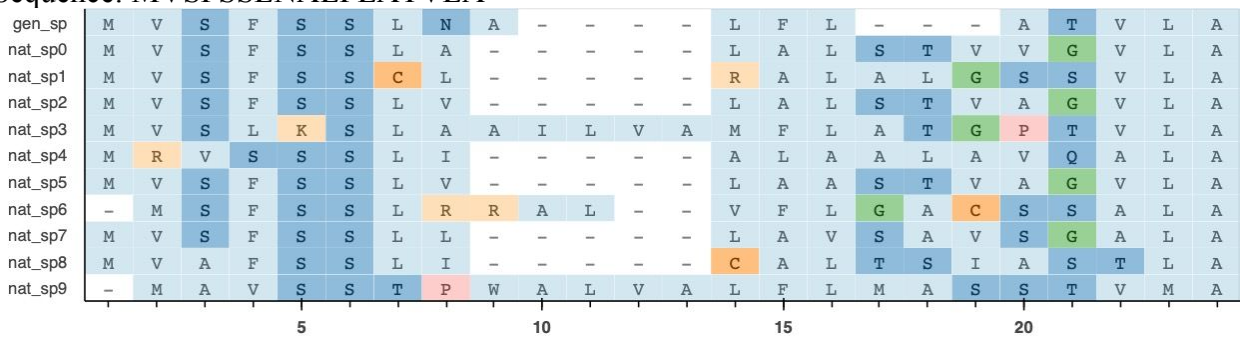
Sequence: MVSFKSALFAAAAVATVADA



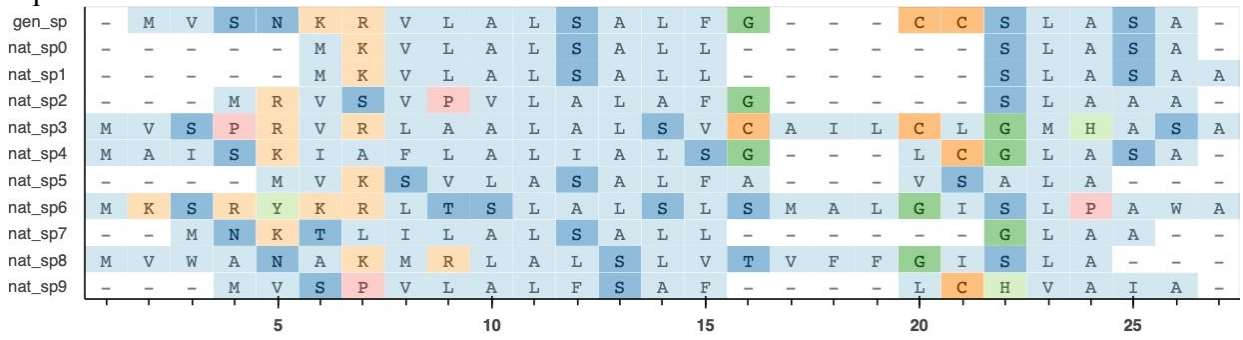
Sequence: MVSFSSLLAAASLAVVNA



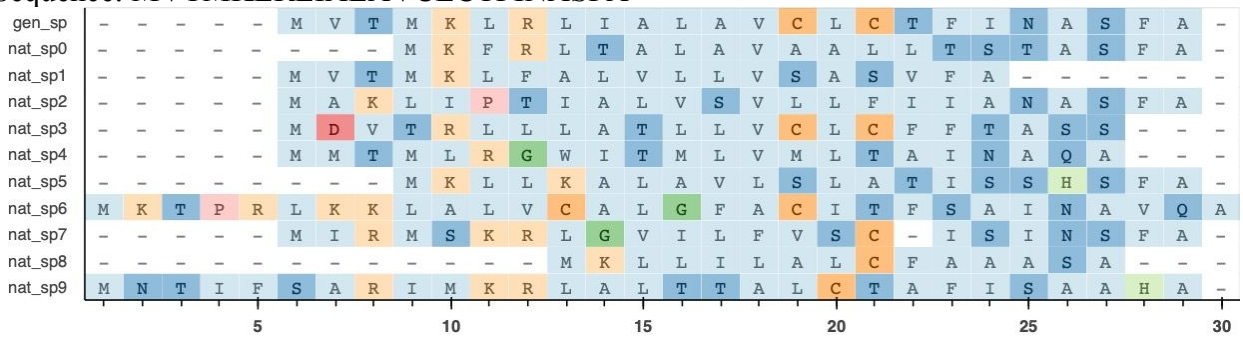
Sequence: MVSFSSLNALFLATVLA



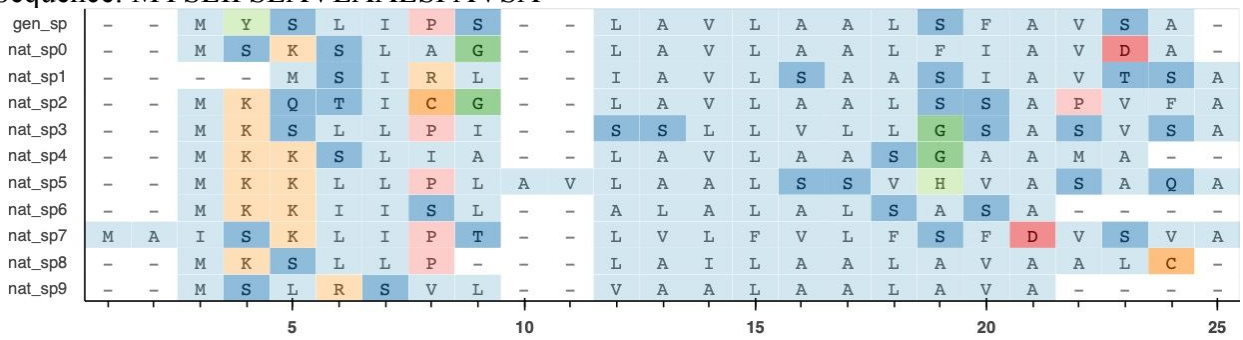
Sequence: MVSNKRVLALSALFGCCSLASA



Sequence: MVTMKLRLIALAVCLCTFINASFA



Sequence: MYSLIPSLAVLAALSFAVSA



Supp Section VI: Bibliography

(1) Kingma, D. P.; Welling, M. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*; 2014.

(2) Madeira, F.; Park, Y. M.; Lee, J.; Buso, N.; Gur, T.; Madhusoodanan, N.; Basutkar, P.; Tivey, A. R. N.; Potter, S. C.; Finn, R. D.; Lopez, R. The EMBL-EBI Search and Sequence Analysis Tools APIs in 2019. *Nucleic Acids Res.* **2019**. https://doi.org/10.1093/nar/gkz268.

(3) Eddy, S. R. Profile Hidden Markov Models. *Bioinformatics*. 1998. https://doi.org/10.1093/bioinformatics/14.9.755.

(4) Owji, H.; Nezafat, N.; Negahdaripour, M.; Hajiebrahimi, A.; Ghasemi, Y. A Comprehensive Review of Signal Peptides: Structure, Roles, and Applications. *Eur. J. Cell Biol.* **2018**, *97* (6), 422–441. https://doi.org/10.1016/j.ejcb.2018.06.003.

(5) Almagro Armenteros, J. J.; Tsirigos, K. D.; Sønderby, C. K.; Petersen, T. N.; Winther, O.; Brunak, S.; von Heijne, G.; Nielsen, H. SignalP 5.0 Improves Signal Peptide Predictions Using Deep Neural Networks. *Nat. Biotechnol.* **2019**, *37* (4), 420–423. https://doi.org/10.1038/s41587-019-0036-z.